LEARNING EMBODIED MODELS OF ACTIONS FROM FIRST PERSON VIDEO

A Dissertation
Presented to
The Academic Faculty

By

Yin Li

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing, College of Computing

Georgia Institute of Technology

December 2017

LEARNING EMBODIED MODELS OF ACTIONS FROM FIRST PERSON VIDEO

Approved by:

Professor James M. Rehg
School of Interactive Computing
Georgia Institute of Technology

Professor Irfan Essa
School of Interactive Computing
Georgia Institute of Technology

Professor James Hays
School of Interactive Computing
Georgia Institute of Technology

Professor Serge Belongie
Department of Computer Science
Cornell University and Cornell Tech

Professor Kristen Grauman
Department of Computer Science
University of Texas at Austin

Date Approved: August 5, 2017

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. James M. Rehg for his continuous support of my Ph.D study and research. His guidance helped me in all aspects of my research and the writing of this disseration.

In addition, I would like to thank the rest of my thesis committee: Prof. Irfan Essa, Prof. James Hayes, Prof. Serge Belongie and Prof. Kristen Grauman, for their insightful comments and all of their hard questions.

My thanks also go to my collaborators, Dr. Piotr Dollár, Prof. Svetlana Lazebnik, Prof. Alan L. Yuille and Prof. Vikas Singh. They have been extremely supportive of my research and my future career.

Finally, many thanks to my family and my friends. Without their precious support it would not be possible to finish my Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Advances in sensor miniaturization, low-power computing, and battery life have enabled the first generation of mainstream wearable cameras. Millions of hours of videos are captured by these devices every year, creating a record of our daily visual experiences at an unprecedented scale. This has created a major opportunity to develop new capabilities and products based on computer vision. Meanwhile, computer vision is at a tipping point. Major progress has been made over the last few years in both visual recognition and 3D reconstruction. The stage is set for a grand challenge that can break our field away from narrowly focused benchmarks in favor of in the wild, long-term, open world problems in visual analytics and embedded sensing.

My dissertation focuses on the automatic analysis of visual data captured from wearable cameras, known as First Person Vision (FPV). My goal is to develop novel embodied representations for first person activity recognition. More specifically, I propose to leverage first person visual cues, including the body motion, hand locations and egocentric gaze for understanding the camera wearer's attention and actions. These cues are naturally "embodied" as they derive from the purposive body movements of the person, and capture the concept of action within its context.

To this end, I have investigated three important aspects of first person actions. First, I led the effort of developing a new FPV dataset of meal preparation tasks. This dataset establishes by far the largest benchmark for FPV action recognition, gaze estimation and hand segmentation. Second, I present a method to estimate egocentric gaze in the context of actions. My work demonstrates for the first time that egocentric gaze can be reliably estimated using only head motion and hand locations, and without the need for object or action cues. Finally, I develop methods that incorporate first person visual cues for recognizing actions in FPV. My work shows that this embodied representation can significantly improve the accuracy of FPV action recognition.

CHAPTER 1

INTRODUCTION

The recent explosion of interest in wearable cameras is fueled by two inter-related factors. First, advances in sensor miniaturization, low-power computing, and battery life have enabled the creation of compact wearable cameras capable of capturing high-resolution video for several hours on a single battery charge. Second, the growth of visual imagery as a key element of social media has created increased demand for convenient solutions for video capture and sharing. The first generation of wearable cameras that are available commercially, ranges from successful consumer products, such as GoPro[1] and Pivothead[2], to experimental prototypes, such as Google Glass[3] and HoloLens[4], and to research platforms, such as Tobii Pro[5] or SMI Eye tracking Glasses[6]. More importantly, people are willing to use these devices to record their daily life, and share their videos online. A recent study [1] showed that $\sim 12\%$ of short videos on social media are captured from the first person perspective. As a consequence, a large corpus of first person videos have been made available over Internet.

Meanwhile, computer vision is poised for a tipping point. Ever since Marr's seminal work in the 1980s [2], our community has postulated the idea of computational vision as an information-processing system. Such a system constructs a representation of the visual world and produces a desired description. We have witnessed major progress in this direction. In particular, the paradigm of "Internet Vision" has enabled the creation of the first generation of large scale vision datasets, e.g. ImageNet [3], which was followed by the

[1] http://www.gopro.com
[2] www.pivothead.com
[3] http://www.google.com/glass/start
[4] http://www.microsoft.com/microsoft-hololens
[5] https://www.tobiipro.com/
[6] https://www.smivision.com/

recent success of deep learning [4]. We can now build deep models with tens of millions of parameters. These models is able to effectively exploit millions of training samples and reliably predict the categorical labels of unseen images. Deep models have also been shown to generalize to a number of vision tasks beyond image classification, including object detection [5], semantic segmentation [6] and action recognition [7].

However, this paradigm considers the visual perception problem in isolation, and ignores the fact that visual perception in humans and animals is coupled with the action of the body. Our perception guides our actions, and our actions in turn influence how we perceive our environment. This perception-control loop allows us to elicit appropriate inputs by conjuring flows of sensory inputs, learning effective visual representations via interactions with our environment and facilitating the performance of complex actions. Moreover, our daily visual experience is vastly different from the visual data on Internet. It thus remains unclear how well models developed on curated datasets collected from Internet will generalize to our daily visual experiences.

The development of wearable cameras and the advance of computer vision make it possible for the first time in history to collect and analyze a large scale record of our daily visual experiences, in the form of first person videos. The analysis of first person videos is called First Person Vision (FPV), also known as Egocentric Vision. FPV allows us to capture human visual experience, infer human body movements, and thus study human actions or activities within the natural contexts in which they occur. Moreover, FPV enables a grand challenge for computer vision to break away from narrowly focused perception problems in favor of in the wild, open world perception-control problems.

1.1 Objective

My thesis work focuses on FPV, i.e. the automatic analysis of videos captured from wearable cameras. My goal is to develop novel **embodied** representations for understanding the camera wearer's actions, by leveraging first person visual cues derived from first person

videos. These cues includes head motion, hand locations and gaze. I consider a visual representation to be "embodied" if it derives from the purposive movements of the camera wearer. There is a distinction between an "embodied" representation and a traditional visual representation, as the "embodied" version aims to couple how we act and what we perceive. Thus, two subjects moving through the same environment would be expected to construct different embodied representations of the same scene, as they might act very differently in the scene and perceive different aspects of the visual environment.

My work investigates three key aspects of first person actions, defined as the person's intentional body movement that are performed to achieve purposeful goals. First, I present a large scale FPV dataset to explore unique properties of first person videos. More importantly, I present first person visual cues, an "embodied" representation of first person actions. Second, by leveraging the proposed representation, I propose a novel model to estimate how the person allocates visual attention when performing a task. Finally, I further incorporate the proposed representation for improving action recognition in FPV.

1.2 Thesis Statement

First person visual cues are unique properties of first person videos, which provide an embodied representation for estimating attention and recognizing actions.

1.3 Overview

What is unique about first person videos? And how does our understanding of FPV help to advance our knowledge of computer vision? Body-worn cameras make it possible to continuously capture human visual experiences in the natural environments in which they occur, and enable the analysis of activities from an embodied vantage point. FPV thus facilitates an embodied approach to perception by placing activities in their natural context. My thesis work exploits a unique property of first person video—the motion of the camera through the scene is fundamentally guided by the intentions and goals of the camera-wearer.

This dissertation begins by describing the construction of a large scale video dataset for understanding first person actions. I then present a set of first person visual cues by estimating the body motion of the person, including head motion, hand movement and gaze shifts. These cues are unique properties of FPV and can be used as an "embodied" representation of FPV actions. First, I will show that these cues are highly correlated, as the person has to coordinate his or her head, hands and eyes to accomplish the task. It is thus possible to infer one set of cues, e.g. egocentric gaze from the rest. Second, I will also demonstrate that the proposed cues contain important information about high level tasks, as these primitives are basic units for executing an action. They can thus be used as complimentary features for improving action recognition, in addition to appearance and motion features.

My research in FPV can thus be organized into three main topics: *FPV datasets and first person visual cues* (Chapter 3), *FPV gaze estimation* (Chapter 4), *FPV action recognition* (Chapter 5). I present an overview of each topic in this section.

1.3.1 FPV Datasets and First Person Visual Cues

The first problem I address is what is unique about first person videos. This question can not be answered without the support of data. And there were no previous video dataset for FPV action recognition, which are comparable to datasets for generic action recognition, such as UCF101 [8] or HMDB [9].

It has been a problem since the start of my dissertation. I first collaborated with Dr. Alireza Fathi to collect our previous dataset of GTEA Gaze+ [10]. I further led the effort of creating and developing the Extended GTEA Gaze+ dataset as an important piece of my thesis work. This dataset subsumes GTEA Gaze+ [10], and include 29 hours of videos from 32 participant, with full set of action annotation, gaze tracking data and annotated hand masks. This is currently the largest dataset in FPV actions and is on par with UCF101 dataset in term of number of samples. We hope our dataset will serve as a major resource

for the research community for a number of core vision tasks in FPV, including action recognition, action detection, gaze estimation and hand segmentation.

By leveraging this dataset, I further demonstrate that there is a set of first person visual cues implicitly embedded in first-person videos. These cues include the camera wearer's head motion and hand location, and they serve as the low level primitives of an action. I further develop computer vision techniques to extract these signals from first person video. Moreover, I build benchmarks for the extracted signal of gaze and hands. Our datasets and analysis provide the first systematic investigation into the unique properties of FPV–first person visual cues. This part is detailed in Chapter 3.

1.3.2 FPV Gaze Estimation

The second problem I address is the estimation of where the camera wearer is looking when performing an action in FPV. Because a person senses the visual world through a series of fixations, egocentric gaze measurements contain important cues about salient objects in the scene, and current actions of the camera-wearer.

I present a model for gaze estimation in egocentric video by using first person visual cues [11]. Specifically, camera wearer's head motion and hand location are combined to estimate where the eyes look. I further model the dynamics of the gaze, in particular fixations, as latent variables to improve the gaze estimation. The proposed gaze estimation results outperform previous algorithms by a large margin on publicly available datasets. In addition, I demonstrate that the predicted egocentric gaze can be used for recognizing daily actions and segmenting foreground objects, leading to a boost in the performance of previous methods in both tasks.

Our work demonstrates for the first time that egocentric gaze can be reliably estimated using only head motion and hand locations derived from first person video, and without the need for object or action information. Moreover, our results on object segmentation and action recognition show that egocentric attention is useful for understanding objects and

actions. This part is described in Chapter 4.

1.3.3 FPV Action Recognition

The third problem I address is the recognition of what the camera wear is doing given a first person video. While action recognition is a well-established topic in vision, first person action recognition remains challenging. For example, significant ego-motion might hamper the motion-based representation that underlie many successful action recognition methods. Moreover, FPV actions face the challenge of "long tailed" distribution—a small set of categories happens frequently, and a huge set occurs with low probability but collectively makes up a critically-important fraction.

I propose to use the proposed first person visual cues for first person action recognition [12]. I show that our set of novel egocentric features can be combined with motion and object features, resulting in a compact representation with superior performance. Moreover, I provide the first systematic evaluation of motion, object and egocentric cues for first person action recognition. This study is performed for both hand-crafted local features and learned deep features. Our benchmark leads to several surprising findings. These findings uncover the best practices for first person action recognition, with a significant performance boost over previous methods on several public datasets.

Our work demonstrates that egocentric cues can be incorporated with motion and object features to improve the performance of egocentric action recognition. Results of our benchmark suggest that the location of egocentric actions is informed by first person visual cues, especially the attention cue. This part is explained in Chapter 5.

1.4 Contributions

My dissertation makes the following contributions:

- I created and developed the Extended GTEA Gaze+ dataset, the largest FPV dataset with gaze tracking data, annotated actions and hand masks during meal preparation

tasks. I establish benchmarks using the proposed dataset for gaze estimation, action recognition and hand segmentation. I believe our dataset and benchmark will provide a major resource for the community.

- I introduce a set of first person visual cues implicitly embedded in first person videos. These cues include the camera wearer's head motion and hand location, and they serve as the low level primitives of attention and actions. They thus forms a novel embodied representation for core vision tasks in FPV.

- I show that for the first time, gaze estimation in FPV is viable with first person visual cues and without an eye tracker. This is done by modeling the coordination of gaze, head motion and hand movement in FPV. By leveraging first person visual cues, I developed a novel method for gaze estimation in FPV.

- I demonstrated that incorporating first person visual cues can significantly improve the performance of FPV action recognition. In particular, I found that the ability to reason about attention is critical for accurate action recognition. This is established via our systematic study of first person visual cues, object cues and motion cues for first person action recognition.

CHAPTER 2
RELATED WORK

In this chapter, I describe the primary background literatures for my thesis work. These previous works are organized into four major parts:

- The section on *Embodied Cognition* provides a perspective from psychology and cognitive science on the coordination of visual perception and body motion in daily life actions.
- The section on *Action Recognition* reviews the vision literature prior to the advent of first person vision
- The section on *First Person Vision* provides a survey of FPV problems and approaches and positions my work in context.
- The section on *Visual Prediction* describes recent efforts to model the long-term temporal relationship between vision and action.

2.1 Embodied Cognition

Embodied cognition is an area of cognitive science and psychology that emphasizes how an agent's cognition is strongly influenced by aspects of an agent's body beyond the brain itself [13]. This is in contrast to the traditional cognitive theories focusing on mental representations in abstraction from bodily mechanisms of sensory processing and motor control [14]. Perhaps the most extreme version of embodied cognition is given by Gibson [15]. He made a strong argument that the goal of visual perception is to learn high order invariants from sensory input and it is only made possible by our body movement. Embodied cognition led further to the development of active vision [16, 17], where the vision system

is actively seeks information that is relevant to current cognitive activity [18, 19]. Among the rich literature in this area, we briefly review the most relevant work on eye, hand and head coordination for actions and activities.

Land and Hayhoe [20] studied gaze behavior in natural tasks such as tea making. They found eye fixation usually precedes hand movement by a fraction of second, indicating eye movements are planned into the motor pattern and lead each action. Ballard et al. [21] investigated functional critical operations occurring at a time scale of one-third second, within a block-copying task. They claimed that at this level the body movements can be matched to the decision systems through implicit references, called deictic, in which pointing movements are used to bind objects in the world to cognitive programs. Pelz et al. [22] explored longer-term temporal coordination of eye, head, and hand movements while subjects performed a similar block-copying task. They discovered regular, rhythmic patterns of eye, head, and hand movements, and argued that these patterns are set by global task constraints. The fact that we need to coordinate our body movements to perform actions in natural environment, has inspired my work on gaze prediction and action recognition.

More recently, Yu et al. [23] studied "embodied intentions", i.e. the use of eye and body movement at the early stage of lexical acquisition. In their study, a teacher is presenting stories using a foreign language. Adult participants are exposed to three sets of learning materials, including first person video with audio (intention-cued), third person video with audio (audio-visual) and audios only. The group in the intention-cued condition performed the best at learning visually grounded meanings for verbal words. It is hypothesized that the subjects used teacher's eye and body movements to track and isolate salient aspects of the scene. They further demonstrate that a computational learning model using a neural network benefits from embodied cues. My dissertation advances this idea by supplementing visual features with first person visual cues for action recognition.

The works that are the most relevant to this dissertation are [24, 25]. Yu and Ballard [24] proposed to recognize first person actions in an office setting by parsing eye movement data

and motion sensor data of hands and head positions. The actions recognized involve only one object, are recorded in a controlled environment and are typically isolated. This work is further extended by Yi and Ballard [25]. They considered more complex behaviors, such as making a sandwich, which consist of a series of actions and involving multiple objects. The sequencing of the actions is modeled by dynamic Bayes network, which is in turn used for recognition. These behaviors are again captured in a highly controlled setting. My dissertation further extends the core idea of this line of research by (1) using a single camera without motion sensors and thus pursuing the development of vision methods; (2) modeling head, eye, hand coordinations in complex, naturalistic actions and activities.

2.2 Action Recognition

Before addressing the connection between action and perception in the context of first person vision, we briefly review the literature on action recognition in a conventional computer vision context. The main property of this work is that it assumes a third person view of one or more individuals, e.g. as would be captured by a surveillance camera, and asks what the individuals are doing. A thorough survey of this previous work is beyond our scope, and we refer the readers to recent review papers [26, 27] for a comprehensive description. In this section, we focus on relevant work on action recognition using body pose, appearance features and mid-level features.

Early works on action recognition started by tracking and classifying articulated body motions [28, 29]. This idea of first tracking the body pose of the actor, and then recognizing the actions based on the trajectory of poses is straightforward. However, it underestimates the challenge of human body pose estimation [30]. More recent approaches leverage learning methods to tightly couple pose estimation and action recognition [31, 32, 33]. Lv and Nevatia [31] proposed to directly match poses against a large set of rendered synthetic 2D poses of desired actions. Wang et al. [32] started with top K-best estimation of body poses, and incorporate segmentation cues and temporal constraints to select the best poses for ac-

tion recognition. Yao et al. [33] modeled the appearance and motion of body parts as the nodes of graph, and represent an action as a mixture of pose templates encoded by an And-Or graph. Action detection was thus done by template matching, using an efficient graph inference algorithm. Note that we also tap features related to body pose in inferring head and hand movements when analyzing first person videos. As we will show in Chapter 3, the first person perspective provides a unique direction for such analysis.

Given the difficulty of recovering accurate and detailed body motion from videos, an alternative approach is to develop more abstract features that encode appearance and motion information. Examples include spatio-temporal interest points and more recent CNN features. Laptev [34] introduced the Space-Time Interest Point (STIP) by extending 2D Harris corner to 3D. Wang et al. [35] proposed to densely sample feature points and track them using optical flow. Multiple descriptors, including HOG [36], HoF [37], MBH [35] or Cuboids [38] can be computed around the interest points, followed by a bag-of-features representation for action recognition. These spatial-temporal descriptors aim to identify key features that are relevant to actions. Several recent work have demonstrate the success of using deep CNN for action recognition [7]. Simonyan and Zisserman [7] proposed the two-stream network that learns to recognize an action from both optical flow fields and video frames. Wang et al. [39] further extend two-stream network to model multiple temporal segments within the video. Du et al. [40] replaced 2D convolution with spatial temporal convolutions and learns a 3D convolutional neural network for action recognition. Carreira and Zisserman further proposed a two-stream 3D convolutional architecture for action recognition [41]. CNN features can be also combined with each tracked local descriptors [42] or human poses estimation [43] to improve the performance. My dissertation builds upon existing deep learning methods and explores how to incorporate first person visual cues into these successful representations in Chapter 5.

There is a growing interest in using mid-level features for action recognition. Yao et al. [44] combined pose estimation and appearance features for action recognition. Fathi

and Mori [45] proposed to construct mid-level motion features from optical flow using AdaBoost. Raptis et al. [46] extracted action parts by forming clusters of trajectories. Recognition was then formulated as matching a subgraph of parts to a template. Tian et al. [47] extended the deformable part model to 3D volumes and learned 3D spatio-temporal parts for action detection. Jain et al. [48] demonstrated that mid-level discriminative patches can be mined from video, and used for classification as well as building correspondences between videos. Most recently, Mathe and Sminchesescu [49] proposed to recognize actions by sampling local descriptors from a predicted saliency map. In Chapter 3 we will argue that egocentric videos contain particularly rich signals about the camera wearer's movements that can be used as mid-level features for understanding the first-person's actions.

2.3 Datasets for Action Recognition

Standardized, large-scale video datasets for human actions has been a major driving force for the advance of action recognition. Examples include UCF101 [8] and HMDB [9] and the more recent Kinetics dataset [41], where tens of thousands of video clips of human actions were collected from Internet and annotated manually. Unfortunately, such a large-scale dataset is missing for FPV action recognition. My early work on the GTEA Gaze+ [10] dataset created the largest benchmark for FPV by far, yet remained an order of magnitude smaller than UCF101. In Chapter 3, I describe my thesis work on the Extended GTEA Gaze+ dataset that attempts to address this gap. Perhaps the most relevant effort to our dataset work is the MPII-Cooking dataset [50]. Both datasets focus on cooking activities, with MPII following a conventional 3rd person paradigm. Our dataset, in contrast, was captured from the first person perspective, and it offers the largest benchmark for FPV action recognition, gaze estimation and hand segmentation. Another highly relevant dataset is the ADL dataset from Pirsiavash and Ramanan [51], where they collected and annotated 10 hours of FPV videos. However, ADL is targeted for complex activities (Activities of Daily Living) and is substantially smaller than our dataset in terms of number of instances.

We provide a detailed comparison of our dataset to previous action recognition datasets in Section 3.3.2 of Chapter 3. We believe that our Extended GTEA Gaze+ dataset can serve as a major resource for the community to further advance the understanding of attention and actions in FPV.

2.4 First Person Vision

We now describe the emerging field of first person vision and its relevant related work. In this section, we focus on predicting gaze and actions in FPV. Other efforts include egocentric hand analysis [52, 53, 54, 55], pose estimation [56, 57], physiological parameter estimation [58, 59], user identification [60, 61, 62] and video summarization [63, 64, 65]. A recent survey of this literature can be found in [66].

FPV Gaze Estimation: My work [11] was among the first to consider gaze estimation in first person videos [67]. We modeled the coordination of eyes, hands and head, and proposed to estimate egocentric gaze using hand and head cues. Our previous work [10] explored a joint model for egocentric gaze and actions. Egocentric gaze is also used to identify important objects [68] or summarize videos [63]. Going beyond egocentric gaze from a single camera wearer, Park et al. [69] proposed to estimate 3D social gaze of multiple persons by identifying regions where the directions of wearable cameras intersect. In addition, Park et al. [70] considered the problem of social gaze by estimating face directions in 2D and projected them into 3D.

FPV gaze estimation is also connected to visual saliency modeling literature [71]. However, these two topics differ in several aspects. FPV gaze is primarily controlled by the task of the camera wearer within a dynamic environment, and thus falls into the category of top-down attention [72]. In contrast, visual saliency modeling usually considers bottom-up attention during the free-viewing of a static image displayed on a monitor [73]. Thus, previous methods for saliency detection produce unsatisfactory results in FPV, as I will show in Chapter 3.

FPV Actions and Activity Recognition: There were a number of early efforts to tackle action recognition in first person videos. Spriggs et al. [74] proposed to segment and recognize daily activities using both first person videos and wearable sensor data. Fathi et al. [75] presented a joint model of objects, actions and activities. Pirsiavash and Ramanan [51] further advocated for a object-centric representation of first person activities. Fathi and Rehg [76] proposed to differentiate egocentric actions by modeling the change of states of objects and materials in the environment. Other efforts included the modeling of conversations [77] and reactions [78] in social interactions.

In these earlier works, several papers [75, 76, 51] reported that local spatial-temporal features (such as [35]) often fire at locations irrelevant to an action due to the camera motion, and thus lead to unsatisfactory performance. In Chapter 5, I present our work on action recognition, initially published in [12], which provides a systematic study of object, motion and egocentric cues, and demonstrated that first person visual cues can significantly improve the performance of FPV action recognition. In particular, we show that motion cues can be exploited effectively within an egocentric framework. Another possible solution is an object-centric representation [51].

Several subsequent works [79, 80, 81] have also identified this critical role of egocentric attention and body movements for understanding first person actions and activities. For example, Kitani et al. [82] encoded optical flow into a global motion descriptor to discover egocentric actions. Su et al. [83] used a similar global motion descriptor for engagement detection. Ryoo and Matthies [84] combined global and local motion cues for interaction recognition. These global motion descriptors provide a coarse representation of the body movement of the camera wearer. In other work, Bambach et al. [80] used hand regions to recognize table-top playing actions. An extension to our work in [12] using deep models was explored by Singh et al. [79], where a multi-stream CNN is learned to capture different type of cues. Similarly, Ma et al. [81] learned a multi-task CNN for egocentric hand segmentation, object localization and action recognition.

2.5 Visual Prediction

There is a growing interest within the computer vision community in visual prediction using videos. The target prediction can vary from a future frame [85] or a representation of a future frame [86], to the goal of an on-going event [87, 88, 89] or a future event [90]

Frame Prediction: Prediction fits naturally into the framework of sequence learning. Sequential models such as Recurrent Neural Networks (RNNs) have thus been used for visual prediction. Ranzato et al. [91] learned a RNN to extrapolate future frames in a video. Similarly, Oh et al. [85] proposed to use a RNN to decode a future frame based on previous frames and control inputs when playing Atari games. A simplified version of sequential models is single step or fixed steps prediction. Villegas et al. [92] combined past motion fields and video frames for next frame prediction.

Visual prediction has recently received considerable attention, as they provide means of self-supervised learning of visual representations for recognition tasks. For example, Vondrick et al. [86] proposed to predict future visual representations in a ConvNet, and reused the learned representation to anticipate future actions. Vondrick et al. [93] further extend their previous work by using a generative adversarial network for frame prediction. Luo et al. [94] proposed to learn a visual representation for activity recognition by predicting a sequence of future flow fields given two frames. Walker et al. [95] estimated future motion of pixels from an image using a variational auto-encoder. Villegas et al. [96] proposed to predict long term future by hierarchical prediction. They consider videos with human actions. And the prediction of future frames is done in two steps. First, they predict the future poses of the actor. Second, they generate future frames conditioned on predicted pose and previous video frames. A similar idea is also explored by Walker et al. [97].

Early Event Recognition and Future Event Prediction: Several works addressed the problem of recognizing the goal of an ongoing events, also known as early event recognition. Schindler and Van Gool [98] presented a system for action recognition from a short

snippet of 1-10 frames. Their results suggested the possibility of recognizing an action given partial observations. Ryoo [87] proposed a dynamic bag-of-words approach for inferring the ongoing the activity from its beginning part. Minh and Da la Torre [88] extended the Structured Output SVM to sequential data as an early event detector. The detector was designed to recognize partial events, as the score of the detector is expected to increase as the event progresses. Li and Fu [99] addressed activity prediction by mining the sequential pattern of action units and building a variable order Markov model.

Taking one step forward, there are only a few work on the challenging problem of predicting a future event. Pei et al. [100] proposed to predict the intent of an activity by parsing video events based on Stochastic Context Sensitive Grammar. Their prediction relied heavily on the scripting of the activities. Xie et al. [101] proposed to infer a pedestrian's intent and predict their trajectories using a probabilistic graphical model. Kris et al. [90] combined the semantic scene labeling with a hidden Markov Decision Process to model the agent's behavior and forecast its trajectory. Vondrick et al. [86] considered the problem of recognizing future actions.

Visual Prediction in FPV: Despite these literatures, there are very few work on visual prediction in FPV. Zhou and Berg [102] proposed to predict the temporal order of two video snippets, as a simplified prediction problem. Park et al. [103] incorporate physical constraints of obstacles and walking avoidance to regress egocentric future trajectories. A concurrent work from Zhang et al. [104] explored the prediction of future gaze in FPV. Predicting future body motion of the camera wearer is a natural extension of my dissertation. And I plan to further explore this topic in the future.

CHAPTER 3

FPV DATASETS AND FIRST PERSON VISUAL CUES

This chapter lays out the foundation of my thesis work. I will start by discussing the key properties and limitations of wearable cameras. I will then describe our effort in constructing video datasets for FPV. These datasets are designed to address key challenges in FPV, including action recognition and gaze estimation. Finally, I will present a set of first person cues that uniquely characterize FPV and can be inferred from first person videos. These cues include head motion, hand location and gaze information, and will be used throughout my dissertation.

3.1 Contributions

I made two major contributions in this chapter.

- I created and developed the Extended GTEA Gaze+ dataset, the largest and most comprehensive FPV dataset to date. Our dataset include FPV videos, gaze tracking, annotated hand masks and actions, and establish benchmarks for FPV gaze estimation, hand segmentation and action recognition.

- I introduce a novel set of first person visual cues, which uniquely characterize first person videos. These cues include the camera wearer's head motion, hand locations and gaze. I also present methods for extracting these cues from videos. Our dataset and study thus provide the first systematic investigation into first person visual cues.

I collaborated with Dr. Alireza Fathi for creating our previous GTEA Gaze+ dataset, the precedent of Extended GTEA Gaze+. The collection of Extended GTEA Gaze+ was in collaboration with Dr. Maithilee Kunda and Mike Lee. In addition, I would like to thank

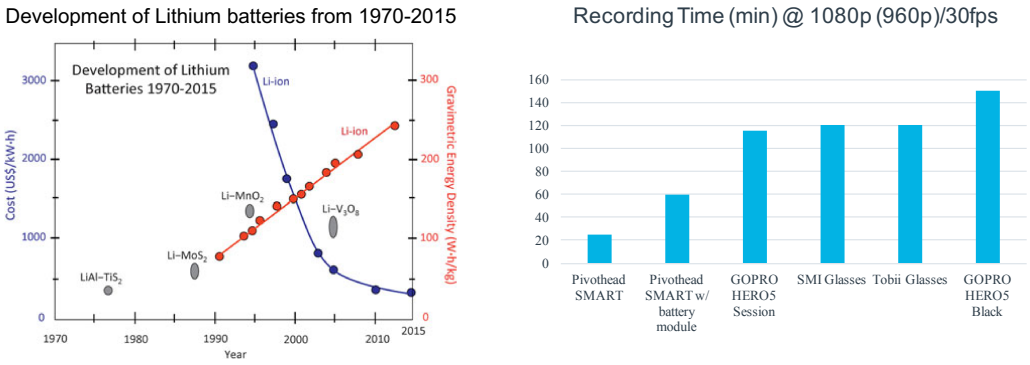

Figure 3.1: A comparison of battery energy density and battery life for wearable camera platforms. Left figure from [105]: the power density of batteries keeps increasing over the last 45 years and their cost decreases. Right: battery life of several commercial platforms. The current limitation is around 2 hours of continuous recording. This is sufficient time to capture most of our daily activities. Figure credit: Yun Zhang.

our data annotation team for annotating our dataset. My thanks also go to Yun Zhang for her help with Figure 3.1 and 3.2.

3.2 Properties of Wearable Cameras

The design parameters for wearable cameras and the mounting of the cameras on the body have a significant impact on the quality of recorded videos. This section discusses several key parameters, including image quality, battery life, field of view and camera mounting.

Image Quality is a basic concern for wearable cameras. Due to recent advance of the imaging sensors, most platforms support the recording of HD videos (1080P or 960P) at more than 24Hz. It is true that many platforms still suffers from traditional issues, such as motion blur, the rolling shutter effect, vignetting and over-exposure. However, these artifacts are less noticeable on modern imaging sensors.

Battery Life is an important practical issue for wearable camera platforms. The battery industry has made steady progress in increasing the power density of lithium-ion batteries over time, while simultaneously driving down their cost, as shown in Figure 3.1 (left) [105]. Standard recording durations on a full charge are shown in Figure 3.1 (right). While the

Sensor	Horizontal FOV	Vertical FOV
Human Binocular	~114°	~100°
Human Combined Monocular	**~200° - 220°**	**~100°**
GoPro	123°	95°
SMI	80°	60°
Tobii	82°	52°
Pivothead	77°	43°

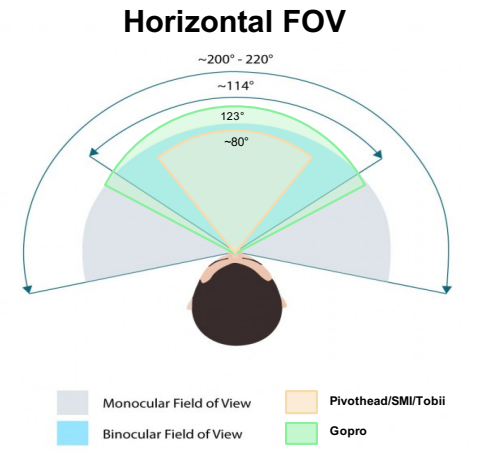

Figure 3.2: A comparison of the Field Of View (FOV) between human vision system and several wearable camera platforms. Left: a list of FOV in both horizontal and vertical directions. Right: an illustration of different FOV in the horizontal direction. GoPro is currently the only platform that offers a FOV on par with our binocular vision. Most head-mounted cameras suffers from a reduced FOV. Figure credit: Yun Zhang.

details vary, the current limit is about two hours of continuous recording of high-quality video, and can be further extended by using external battery packs. This battery life is sufficient to capture most of our daily activities, yet is still far away from being able to record a full day's worth of video on a single charge.

Field Of View (FOV) is currently the major issue for most wearable cameras. FOV determines how much of the world the camera can see at any time. A key property of human vision is the wide FOV. GoPro understood the importance of capturing from a large field of view when recording sports activities. In figure 3.2, the left panel lists a comparison of FOV for several commercial platforms, while the right panel illustrates the horizontal FOV. At 123 degrees in the horizontal, the GoPro camera covers the full binocular range of human stereo depth perception. But by combining both eyes, humans can achieve a remarkable horizontal FOV of more than 200 degrees. In comparison to GoPro, head-mounted cameras, such as Pivothead or SMI, suffer from a reduced field of view. This can be a challenge for FPV as a significant portion of the peripheral visual field is truncated by the reduced FOV.

Camera Mounting is another significant issue. How the camera is mounted determines what it can see. For example, a chest mounted camera will not always capture what the camera wearer is looking at. My work is therefore based on **head-worn cameras**. The head-mounted camera is promising because (1) it offers an additional cue of head motion; (2) it best captures the camera wearer's visual experience; (3) it allows us to easily synchronize the captured video with gaze tracking data; (4) it is supported by a large set of hardware platforms.

In the rest of the document, I will assume first person videos coming from a head-mounted camera that can produce high quality videos with a sufficient FOV to cover important aspects of FPV actions. This simplification allow us to focus on underlie computer vision problems.

3.3 FPV Datasets for Actions

In this section, I describe several major FPV action datasets. Early in my thesis work, I was involved in the construction of GTEA Gaze and GTEA Gaze+ datasets.[1]. Then I led the effort of developing ISTC Wetlab and Extended GTEA Gaze+ datasets. The Extended GTEA Gaze+ dataset is one of the core contributions of my dissertation. At present, it provides the largest and most comprehensive benchmark for FPV action recognition, gaze estimation and hand segmentation.

3.3.1 Previous FPV Datasets

I provide a brief survey of previous FPV datasets for action recognition. Other relevant FPV datasets include EgoAction dataset [82] for action discovery, UT Egocentric dataset [68] for video summarization and GT Social Interaction dataset [77].

UCI ADL consists of 10 hours of videos from 20 subjects. The dataset is designed for understanding Activities of Daily Living (ADL) in FPV, such as "combing hair" or

[1] GTEA stands for Georgia Tech Egocentric Activity

"brushing teeth". The dataset is annotated with dense temporal boundaries of 18 unique activities and bounding boxes of 10 different objects at sparsely sampled frames. The dataset was collected using a chest-mounted GoPro camera, with a resolution of 1280×960 at 30Hz. It contains 364 instance of activities. While these activities are naturalistic and complex, the number of instances is small. Also, a chest mounted used in this dataset might not capture what the camera wearer is seeing, as the direction of sight can be different from the direction of the body.

JPL Interaction is captured by a camera mounted on a robot. The dataset consists of video clips when a human is interacting with the robot. The goal of the dataset is thus to recognize the interaction between the person and the robot. The dataset comes with pre-segmented action clips with a resolution of 320×240 at 30Hz. The dataset has 94 action instances from 7 action categories, with a total duration around 25 minutes. Actions in the dataset are relatively simple, e.g. "hand shake" and only involve one single person.

GTEA consists of 7 types of meal preparation activities (recipes) in a controlled lab setting, where each activity is performed by 4 subjects. Each session last 1 minute. The dataset was collected using a head-mounted GoPro camera. The camera has a wide angle lens and a resolution of 1280×960 at 30Hz. The dataset comes with action annotations of 61 classes with 456 instances. It was first proposed by Fathi et al. [106] and subsequently refined in our previous work [12]. The dataset suffers from two major issues. First, the size of the dataset is small. With a few hundreds of samples, it will be hard for training complex models such as deep networks. Second, the videos are captured in a lab controlled setting, and thus does not elicit naturalistic actions.

GTEA Gaze includes 17 sequences of meal preparation activities in a lab setting, performed by 14 subjects. Each session last 4 minutes on average. The video was captured by Tobii eye-tracking glasses[2], with a resolution of 640×320 at 30Hz. The dataset comes with action annotations, as well as gaze tracking data. The size of the dataset is small, with

[2] `http://www.tobiipro.com`

25 action classes and 270 action samples. The dataset was first introduced in our previous work [10] and subsequently refined in [12]. The dataset is designed for action recognition and gaze prediction, yet suffers the same issues as GTEA. We have thus proposed GTEA Gaze+ as a replacement of GTEA Gaze.

GTEA Gaze+ is our dataset for FPV gaze estimation and action recognition. GTEA Gaze+ contains 37 video sessions from 6 subjects performing a set of 7 meal preparation activities in Georgia Tech's AwareHome. The dataset thus contains approximately 9 hours of videos captured in an instrumented house with a kitchen that contains all of the standard appliances and furnishings. We used SMI eye-tracking glasses to record this dataset. Each session is captured by a first person video using head-mounted camera with gaze tracking. The video has a resolution of 1280×960 at 24Hz and the gaze tracking is at 30Hz. The dataset first introduced in our previous work [10] and refined in our subsequent work [12].

We have annotated the onset and offset of all actions in each video. Our taxonomy of action names consist of a verb and a set of nouns. such as "put turkey (on) bread". In this case, the verb defines the action the person is performing and the nouns describes the objects that are involved in the action. Note that "take bread" is different from "take tomato" in our taxonomy, although they have the same type of motion "take". Thus, our taxonomy is highly flexible and can capture fine-grained FPV actions. Moreover, we allow the actions to overlap with each other in time. We used the same taxonomy for the rest of our dataset, including ISTC Wetlab and Extended GTEA Gaze+. GTEA Gaze+ is designed for first person gaze estimation, action recognition and activity recognition. The dataset covers a rich set of object manipulation tasks in a naturalistic setting, and contains significant more number of samples in comparison to previous datasets. It is public available at our website `http://cbi.gatech.edu/fpv`.

ISTC Wetlab captures structured activities in a controlled lab setting. The dataset targets a specialized yet important subset of actions in biochemical benchtop experiments. We capture our dataset during standard training protocols for acquiring basic skills in bechtop

experiments. The actions are complex and naturalistic, yet remain structured as they must follow experiment protocols. Thus, the dataset provides a vehicle for studying skilled actions in a highly specialized domain. We believe this effort also offers a new application domain for looking into egocentric actions.[3]

We collected videos and gaze tracking data from 7 subjects, each performing a set of 3 activities, including preparation of media and cultures, manipulation of small volumes and manipulation of large volumes. The data was recorded using using SMI eye tracking glasses[4]. The video has a resolution of 1280×960 with 24 frames per second and the gaze was tracked at 30Hz. Sessions with low gaze tracking quality are discarded, leading to a total number of 20 sessions with an average length of 12 minutes. We follow the same taxonomy as GTEA Gaze+ for annotating the onset and offset of all actions in ISTC Wetlab. Each session contains an average of around 100 action instances with over 60 total different action categories. The dataset is public available at our website `http://cbi.gatech.edu/fpv`.

3.3.2 Extended GTEA Gaze+ Dataset

Extended GTEA Gaze+ is our latest effort for FPV gaze and actions. The dataset was captured using the same setting as GTEA Gaze+, yet at a significantly larger scale. The dataset contains 29 hours of videos from 86 unique sessions of 32 subjects performing 7 different meal preparation tasks. Similar to GTEA Gaze+, the video has a resolution of 1280×960 at 24Hz and gaze tracking is performed 30Hz. Our annotations include ~ 11K actions instances and 15K hand masks. We now describe the motivation of the dataset, as well as how we collect and annotate the data.

[3]ISTC Wetlab dataset is collected at University of Washington, under the support of Intel Science and Technology Center for Pervasive Computing (ISTC-PC).

[4]`http://www.eyetracking-glasses.com`

Why Do We Need Another FPV Dataset?

Large-scale video datasets and their benchmarks have been the main driving force for action recognition. For example, UCF101 provides 13K video clips of human actions. Since the introduction of the dataset in 2012, UCF101 has been used as a major testbed for generic action recognition. And the accuracy of recognition has improved from 70% to 98%. This significant performance improvement reflects our progress on action recognition. However, such a dataset is missing in FPV. The largest FPV action dataset is our GTEA Gaze+, and has less than 2K action samples. The size of the dataset is particularly problematic for modern methods that are data savvy, such as deep models. This had thus motivated us to develop a new dataset–Extend GTEA Gaze+ dataset. We hope our dataset can bridge the gap and provide an interesting opportunity for studying gaze, hands and actions in FPV.

Our new dataset subsumes GTEA Gaze+ as a subset, yet with revised annotations via our new annotation pipeline. By the time of this document, it is the largest dataset for FPV action, gaze and hands. Sample frames of the dataset are shown in Figure 3.3, and compared to our previous datasets. This new dataset contains a rich set of actions under different lighting conditions and with diverse set of objects. All data will be made public available along with my dissertation [5].

Data Collection

The dataset was collected at the kitchen area of AwareHome on Georgia Tech campus. The kitchen area of AwareHome, where videos are recorded, provide a naturalistic household environment that contains the standard appliances, furnishings and food. We have recruited the participants from the student pool of Georgia Tech. To protect the privacy of our participant, we require a written consent for each subject before the sessions. And we have manually post-screened all recorded videos to black out any frames that might reveal the identity of the participant. These frames are usually around the start and end of

[5] Will be available on our dataset website: `http://cbi.gatech.edu/fpv`

Figure 3.3: Sample frames from the videos in Extended GTEA Gaze+ dataset. Our dataset contains videos with different lighting condition, object instances and actions.

a session, with rare cases during a session due to reflections. We maintain $> 95\%$ of all frames after screening.

During a session, the subject was first given a few minutes to get familiarize with the target recipe. The subject was then asked to prepare the dish by following the loosely defined recipe. Our recipe only include key steps of the dishes, such as "boil the pasta", and thus no detail instructions are given. Copies of recipes are available on site as a subject might check them during a session. A subject is not required to cover all steps or the order of the steps in a recipe as long as they can finish the dish. We have also deliberately altered the lighting condition and the object instances in the kitchen from session to session. Thus, the videos exhibit a high diversity in terms of lighting, objects and actions.

Data Annotation

We have supplement the datasets with two types of annotations: frame-level annotation of actions for all videos, and pixel-level annotation of hands for sparsely sampled frames. In this section, we describe our annotation pipelines for both tasks.

Action Annotation: Our previous datasets, such as GTEA Gaze, were annotated us-

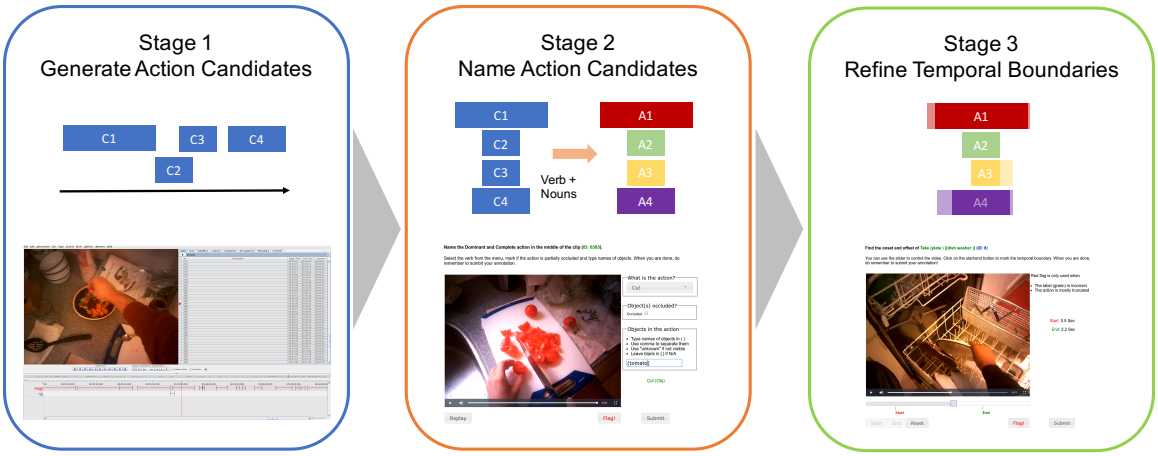

Figure 3.4: Action annotation pipeline. We follow a three stage pipeline for annotation, with each stage focuses on a single task. From left to right: ELAN interface for action candidates; Web UI for action naming; Web UI for action trimming. We developed the web UI based on [108] and made it public available at `https://github.com/happyharrycn/vatic_fpv`

ing ELAN [107]—A multi-modal annotation tool developed at Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands[6]. We propose to streamline this single annotation task by dividing it into multiple stages. Our new pipeline consists of three main stages, with each stage focusing on a single task. These stages are *action candidate labeling*, *action naming* and *action trimming*. Figure 3.4 include an illustration of our interfaces for all stages. We describe each stage in details.

Action Candidate Labeling aims to identify all potential actions and their rough temporal extents. We use ELAN [107] for the annotation and allow two action candidates to overlap in time. Annotators are ask to mark rough onset and offset of all possible actions in screened videos. Note that we allow a small amount of noise in this stage in trade of efficiency, as the later stages can filter out bad candidates.

Action Naming seek to label the actions using our taxonomy. We first pad all candidates with extra frames at the start and the end, and then crop them into individual clips. These clips are most likely to include the full extent of a single action. We then use a web interface for naming the clips. We present one clip at a time and ask annotators to name it. Cropping

[6] `http://tla.mpi.nl/tools/tla-tools/elan/`

the videos not only reduces the visual content that an user has to examine, but also makes sure that the name of the action can be inferred using the isolated snippet. Moreover, we allow the users to red flag a clip if (1) no actions are presented; or (2) there are multiple major actions; (3) the action is not complete in the given clip.

Each action label includes a single verb from a list of 20 pre-defined dictionary, and a set of nouns from free-form user inputs. These verbs and nouns form atomic units of actions. The verb describes the motion, such as "take", "turn on", and the nouns specify the objects involved in the action, such as "tomato" or "peanut butter container". We did not distinguish between the plural or singular form of the nouns. The combination of verb and nouns can describe complex actions, such as "pour egg mixture (from) bowl to pan". A similar naming taxonomy is also explored in [109] and discussed in [110].

Action Trimming further refines the temporal boundary of named action clips. We present both a video clip (temporally padded) and its label (from previous stage) to annotators, and ask them to identify the temporal boundary of the action. The user can mark the onset and offset using a similar web UI in previous step (see Figure 3.4 for an example). Again, we allow the user to red flag a action clip if its label is incorrect.

We post-process all annotations by following steps to eliminate unreliable samples. For all stages, we ignore actions that are less than 0.5 seconds, as (1) the frames are typically motion blurred; (2) we found it hard to accurately identify their temporal boundaries due to rapid motion. After the second stage of the annotation, we chunk some of the object categories. Specifically, fork, knife and spoon are merged into a single category of "utensil". Moreover, we also prune less frequent actions.

We have found that this new pipeline helps to improved the overall efficiency by at least 50%. Moreover, this multi-stage process allow us to recover from an error in a previous stage. For example, an incorrect action candidate can be reject during labeling. An incorrect action label can be filtered out when trimming the video. All annotations are performed in-house by our trained annotators to ensure the best quality, although the pipeline

Figure 3.5: Ground truth hand masks from our Hand14K dataset. The annotated masks are shown as green regions superimposed on the original frames. Our dataset includes egocentric hands with different poses and lighting conditions.

and tools can be easily crowd-sourced.

At the time of writing this document, all videos had gone through the first two stages of annotation. We obtained 17K action candidates after stage one, 15K named action clips after stage two, and more than 10K action instances after post-processing. These clips have an average duration of 3.1 seconds. The number of action clips is thus on par with UCF 101 (13K), the current major dataset for generic action recognition. Our dataset does have shorter video clip than UCF101 (3.1 vs. 7.2 seconds), as actions in cooking activities tend to happen faster.

Hand masks: Egocentric hands are extremely important cues for object manipulation tasks. To facilitate the analysis of hands, we have annotated 13,847 images with pixel-level hand masks. These images are sparsely sampled frames from all 86 videos in the dataset. We out-sourced the annotation task by using a modified interface from [111]. Each hand mask is thus represented as one or more polygons (if it is been occluded). We obtained a total of number of 15176 hand masks with a average of 1.1 masks per image. These images and hand masks are captured by our Hand14K dataset, an important part of the Extended GTEA Gaze+ dataset. All data will be made available at `http://cbi.gatech.edu/fpv`. Figure 3.5 show sample annotations of hand masks from Hand14k. We hope this dataset will provide a major resource for analyzing hands in FPV.

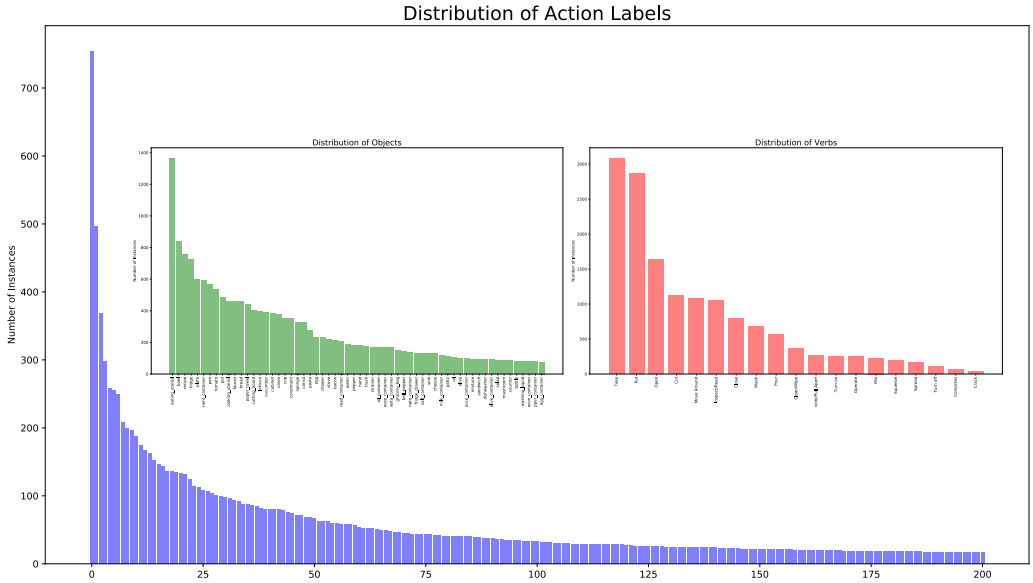

Figure 3.6: The long tailed distribution of verbs, nouns and actions in our proposed Extended GTEA Gaze+ dataset. This is a unique property that characterizes our daily visual experience. Our dataset thus poses the challenge of learning from unbalanced samples.

Statistics of the Dataset

Our final dataset includes 29 hours of FPV videos with a resolution of 1280×920 at 24Hz. The dataset has 86 unique sessions from 32 subjects. Each session consists of a HD video (1280×960), an audio sampled at $44K$Hz, binocular gaze tracking data (30Hz), frame-level action annotations and hand masks at sparsely sampled frames.

Our annotations include 15K hand masks and more than 10K (10119) action instances from 108 action categories. That is around 1 million action frames (out of more than 2.5 millions of frames) with more than 1 millions of tracked gaze points. In particular, our action instances have an average duration of 3.1 seconds with an average of 11 events per minutes. Our hand annotation has an average of 1.1 masks per image. Sample video frames and hand masks can be found in Figure 3.3 and 3.5, respectively.

Table 3.1 shows all our verbs, as well as top 10 object and action labels. The distribution of these labels are illustrated in Figure 3.6. After post-processing, our action annotation includes 19 verbs, more than 50 nouns and 108 unique action labels (a combination

Table 3.1: Annotated action labels for our Extended GTEA Gaze+ Dataset. After pruning less frequent labels, we have 19 unique verbs, more than 50 nouns and a combination of 108 action labels. We show all verbs and top 10 object and action labels.

Verbs (all 19)	Open, Close, Take, Put/Place, Turn on, Turn off, Wash, Cut, Move Around, Operate[7], Pour, Squeeze, Spread, Mix, Crack, Inspect/Read, Compress, Clean/Wipe, Divide/Pull Apart
Nouns (top 10)	Eating Utensil (Knife/Fork/Spoon), Bowl, Recipe, Fridge, Plate, Condiment Container, Pan, Tomato, Pot, Cooking Utensil (e.g. Spatula)
Action Labels (top 10)	Read Recipe, Open Fridge, Take Eating Utensil, Put Eating Utensil, Cut Tomato, Turn on Faucet, Open Cabinet Cut Cucumber, Operate Stove, Close Fridge

Table 3.2: Comparison between FPV datasets. Extended GTEA Gaze+ is the largest egocentric action datasets in terms of the number of subjects, duration, number of action categories and number of instances. Our dataset also provide the most comprehensive benchmarks on action recognition, hand segmentation and gaze tracking.

	UT Ego	EgoAction	JPL Interaction	UCI ADL	GTEA	GTEA Gaze	**GTEA Gaze+**	**ISTC Wetlab**	**E-GTEA Gaze+**
Task	Summary	Action Discovery	Activity	Activity	Action	Action	Action	Action	Action
Mounting	Head	Head	Chest	Chest	Head	Head	Head	Head	Head
Resolution	480*320	840*480	320*240	1280*960	1280*960	640*480	1280*960	1280*960	1280*960
FPS(Hz)	15	30	30	30	30	30	24	24	24
Duration (hours)	20	0.7	0.4	10	0.6	1	9	4	**29**
# Subjects	4	N/A	N/A	20	4	14	6	7	**32**
# Action Categories	N/A	N/A	7	18	61	25	44	51	108
# Action Instances	N/A	N/A	94	364	456	270	1958	2067	**¿10K**
Other Sensors	N/A	N/A	N/A	N/A	N/A	Gaze	Gaze	Gaze	Gaze

of verb and nouns). There is a high diversity in terms of object and action labels, yet the most frequent actions tend to be simple, such as "read recipe" or "open fridge". Figure 3.6 demonstrate the a "long tail" distribution of our action categories. While the most common action "read recipe" happens more than 700 times, the least common action "Pour milk from milk container into bowl" only occurs 30 times. This distribution, which we believe have characterized our visual experience, is very different from all previous action recognition datasets, such as UCF101 or HMDB. This is a significant challenge for action recognition, as the recognition method has to learn from unbalanced samples.

I further compare theses statistics to other action recognition dataset. Table 3.3.2 com-

Table 3.3: Comparison of action recognition datasets. We compare the statistics our Extended GTEA Gaze+ to other generic action recognition datasets, including HMDB [9], UCF101 [8], MPII-Cooking [50], ActivityNet [112], Charades [109] and Kinetics [41]. Our dataset is the largest egocentric action datasets. The size of this dataset is on par with UCF101, the current major benchmark for generic action recognition. Our dataset not only offers gaze tracking data and annotated hand masks, but also facilitates the task of action temporal localization. In addition, our dataset faces the unique challenge of unbalanced action samples.

	HMDB	UCF101	MPII-Cooking	ActivityNet	Charades	Kinetics	E-GTEA Gaze+
Task	Actions	Actions	Actions	Activities	Activities	Actions	Actions
Temporal Localization	No	No	Yes	Yes	Yes	No	Yes
Resolution	320*240	320*240	1624*1224	Varies	Varies	Varies	1280*960
FPS (Hz)	25	25	29.4	Varies	Varies	Varies	24
Duration (hours)	∼10	27	8	648	83	833	29
# Action Categories	51	101	78	200	157	400	120
# Action Instances	7K	13K	13K	20K	67K	300K	11K

pares the proposed Extended GTEA Gaze+ dataset with other FPV action datasets. Our dataset provides the largest benchmark for gaze tracking, hand segmentation and action recognition in FPV. Specifically, the number of action instances in our dataset is a magnitude larger than previous largest dataset. Our dataset thus offers an interesting opportunity for data savvy methods, such as deep networks.

Table 3.3 compares our dataset with other generic action recognition datasets. The size of our dataset is on par with UCF101 [8] and HMDB [9], the current major benchmarks for generic action recognition. However, our dataset offers several advantages than other datasets: (1) it includes HD video with much higher resolution (960P vs. 240P in UCF101); (2) it provides action annotations in video sequences and thus facilitates the task of action detection; (3) it offers a diverse set of fine-grained action labels, including 19 unique verbs and more than 50 unique nouns (objects), leading to a combination of 108 actions; (4) it comes with gaze tracking data for all videos; (5) it elicits the major challenge of long-tailed distribution of actions, which is not available in previous curated datasets.

3.4 First Person Visual Cues

By leveraging the proposed dataset, I present a set of fist person visual cues in this section. These cues capture the body movement of the camera wearer, and can be divided into three parts on head motion, hand locations and gaze points. I show how to extract these cues in FPV. They will be used in later chapters as an "embodied" representation for gaze estimation and action recognition.

3.4.1 Egocentric Head Motion

My dissertation focuses on head mounted cameras. As the camera is aligned with the first-person's head direction, the camera motion is a proxy of the camera wearer's head motion. This motion, albeit simple, encodes non-trivial information of first person's actions. Estimating camera motion is a well-studied topic in computer vision. However, reliable ego-motion estimation under rapid camera motion remains challenging.

Our previous work explored the tracking of sparse interest points [12]. However, sparsely tracked points fail to capture the details of the motion field, which is important for action recognition. In the meanwhile, dense optical flow methods have demonstrate promising results. In addition, dense flow field provides pixel-level motion important for actions and naturally encodes camera motion. Throughout my dissertation, I compute the dense optical flow for building motion representations. To estimate camera motion from the flow field, we keep all pixels, which have high Harris responses and are outside hand masks. We then fit a homography matrix for head motion using RANSAC [113]. A rapid motion is detected if the number of tracked points is small. We can take the translational component of a homography whenever a 2D motion vector is needed, e.g. for gaze estimation.

A major challenge for estimating optical flow in FPV is pixels with large displacements. This happens more often in FPV due to constant ego-motion. Figure 3.7 shows such an ex-

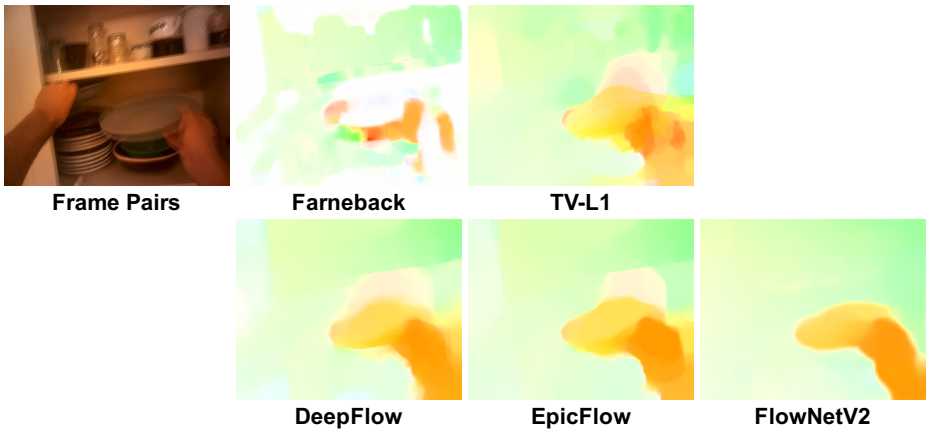

Figure 3.7: Comparison of flow results on a sample pair of frames from our dataset. From left to right: Mean of the frame pairs, Farneback Flow [114], TV-L1 Flow [115], DeepFlow [116], EpicFlow [117] and FlowNetV2 [118]. All flow maps are colored using Middlebury format. We empirically found that FlowNetV2 can consistently better results than other methods.

ample frame pair from our dataset. Here the head motion produces a moderate background motion for all pixels and action of "take" imposes a separate foreground motion. We compare flow results from several widely-used methods. Among them, Farneback [114] and TV-L1 flow [115] are most widely used for action recognition, yet they produce less reliable motion estimate. DeepFlow [116], EpicFlow [117] and FlowNetV2 [118] are more recent methods that achieved good performance on several benchmarks. FlowNetV2 gives the best result in this example.

We have empirically observed the new FlowNetV2 [118] can consistently produce satisfactory flow results on FPV videos. This is probably due to the design of FlowNetV2. The method seeks to gradually refine the flow field using a cascade of deep models. This refine scheme is better at capturing and compensating for global motion induced by the movement of the first person. Practically, the runtime of FlownetV2 on a GPU is only slightly more expensive TV-L1 flow. Therefore, we use the FlowNetV2 [118] for the rest of the work, unless otherwise specified.

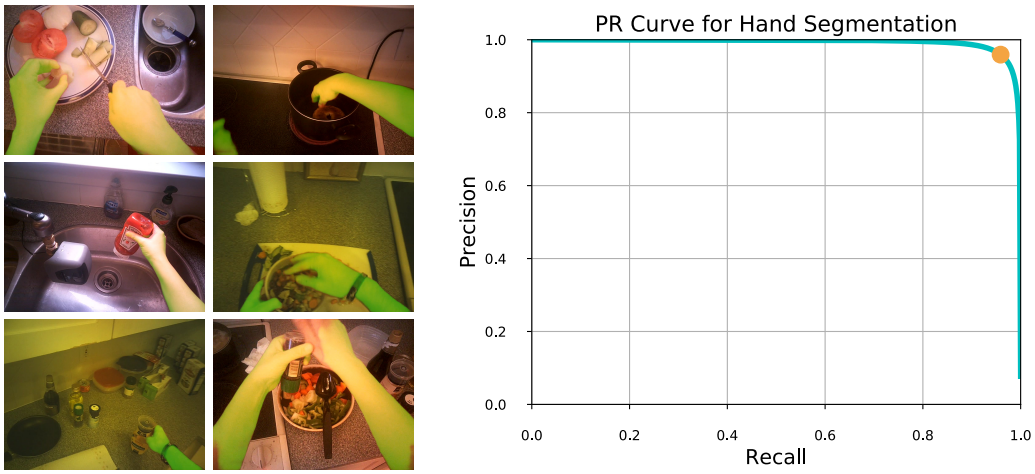

Figure 3.8: Hand segmentation results using fully convolutional network [6]. We show visualization of the segmented hand masks (as the green region) on the left and the precision-recall curve of our test set on the right. While the results are not prefect, e.g. missing hand pixels that are under motion blur or around the boundary of a frame, our method can produce highly accurate hand masks.

3.4.2 Egocentric Hand Cues

Egocentric hands provide an important cue for first person actions. For example, the 2D position of the hand indicates its relative position to the camera—the first person's head in our setting. The shape of the hand induces its pose, and the motion trajectory of the hand tells us how we interact with the objects over time. The information regarding egocentric hands, including pose, location and movement is directly linked to the first person's interaction with objects. We explored different ways of encoding egocentric hands for gaze estimation and action recognition. While their detail will be covered in later chapters, they all require the segmentation of hands. Here we describe our hand segmentation system.

Hand Segmentation: Accurate segmenting and tracking of hands in egocentric video remains a non-trivial problem [52]. We explored using semantic segmentation pipelines [119, 6] for segmentation egocentric hands for all frames in the video. With the help of our new Hand14K dataset and modern deep models, we were able to obtain accurate hand segmentation results. Not surprisingly, we found that deep models performance much better than hand-crafted features, such as TextonBoost [119].

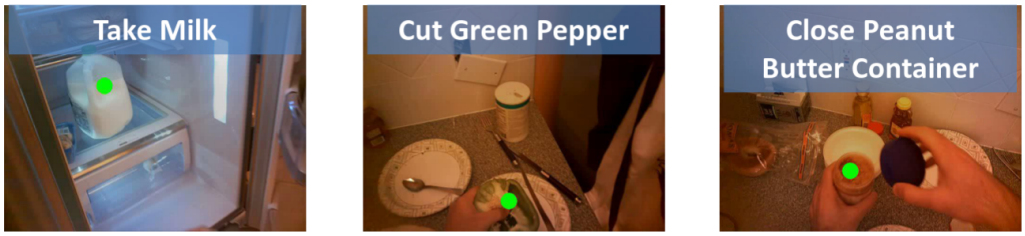

Figure 3.9: Egocentric gaze (green dots) over the video frames tracked using SMI glasses. We also show the current action label. The figure demonstrate that egocentric gaze points often focus on objects and regions that are relevant to current action.

More specifically, we used a modified version of the fully convolutional network [6]. Instead of progressively upsampling and merging the confidence map, we upsample all intermediate maps to the input resolution and linearly combine them into a final hand confidence map. This is similar to the architecture in [120], yet we only use a single loss function at the end. For training, we start with ImageNet pre-trained VGG network [121]. To benchmark our method, we use a fixed split on our new Hand14K dataset. This split has 90% of the frames from 30 subjects for training and 10% frames from two hold-out subjects for testing. We obtained a F1 score of 95.2% using [6]. Figure 3.8 shows visualizations of hand segmentation results and the precision-recall curve on our test set.

3.4.3 Egocentric Gaze Cues

As we sense the visual world through a series of fixations, our gaze reveals important information about our goals. Gaze points often lie on objects that are relevant to the task we are performing, since gaze is used to coordinate actions [122]. In FPV, gaze point is represented by a 2D image point in each frame (see Figure 3.9).

First person's gaze can be directly measured using eye tracking. We use a commercial wearable eye tracker (SMI eye-tracking glasses) to obtain gaze points. The system projects inferred light over the camera wearer's cornea and estimates the gaze direction using glint, i.e. the reflection of the light source on cornea. The system can reach an tracking accuracy of 1 degree with personal calibration. Yet it requires the device to remain fixed on one's

head. We found a higher error for eye tracking during our experiments, mainly due to small motion of the device over the face after the calibration. Regardless of the error, we consider these gaze points as reliable measurements, and use them as ground truth for gaze estimation and action recognition. We also present a method to estimate first person gaze from the video in Chapter 4, when the eye tracker is not available.

3.5 Conclusion

In this chapter, I describe our effort for developing video datasets for FPV gaze and actions. The proposed dataset, called Extended GTEA Gaze+ contains 11K action clips from 29 hours of videos, 15K annotated hand masks sparsely sampled from the video frames, and gaze tracking data for over 2 million video frame. This is by far the largest benchmark for FPV action recognition. My thesis work thus provide the first large-scale dataset for a number of vision tasks in FPV.

More importantly, I present a set of first person visual cues from first person video, including egocentric hand, head and gaze cues. These cues are implicitly embedded in the video, capturing body movement of the first person and revealing the underlie goal and intention of acting. I have described novel egocentric cues, as well as methods to extract them from the video. The dataset and study provide the first systematic investigation into first person visual cues.

CHAPTER 4
FIRST-PERSON GAZE ESTIMATION

Egocentric gaze is defined as the line of sight in a head-centered coordinate system. It is usually represented as 2D points in a video. Estimating first person's gaze or egocentric gaze is a key component in FPV [123]. Because a person senses the visual world through a series of fixations, egocentric gaze measurements contain important cues regarding the most salient objects in the scene, and the actions of the camera-wearer. Previous works have demonstrated the utility of gaze measurements in object discovery [124] and action recognition [10].

This chapter addresses the problem of first person gaze prediction, which is the task of estimating the user's point-of-gaze given an egocentric video. Previous work on gaze prediction in computer vision has primarily focused on saliency detection [125, 126, 127, 71]. Previous gaze prediction models, also known as visual saliency, can be roughly categorized into either (1) bottom-up approaches where the gaze is attracted by the discontinuities of low level features, such as color, contrast and edge; or (2) top-down approaches where the gaze is directed by high level semantics, such as tasks, objects or scene. However, none of these approaches are sufficient to estimate egocentric gaze in the context of hand-eye coordination tasks. Saliency detection can be effective for visual search, but does not identify the key regions in a manipulation task. Task-driven methods can be effective, but require the identification of current activity, which is an open problem in itself. We explore a third alternative: We address the question of whether measurements of head and hand movements can be used to estimate gaze, without reference to saliency or activity models. To simplify the problem, we limit the scope of our method to object manipulation tasks during meal preparation.

During object manipulation tasks, eye, head and hand are in continual motion, and the

coordination of these movements is requisite [22]. For example, large head movement is almost always accompanied by a large gaze shift [128]. Also, the gaze point tends to fall on the object that is currently being manipulated by the first person [128]. These examples suggest that we can model the gaze of the first person by exploring the coordination of eye, hands and head, using egocentric cues alone.

4.1 Contributions

This chapter has three major contributions.

- I demonstrated that for the first time, egocentric gaze in FPV can be reliable estimated by using egocentric head and hand cues, and without using an eye tracker. This is done by leveraging our proposed first person visual cues and without the use of object or action information.

- I present a learning based model for FPV gaze estimation. Our model estimate gaze at each frame, and capture the dynamic of gaze shifts over time. Our method outperforms all previous bottom-up and top-down saliency detection algorithms by a large margin on publicly available datasets.

- I further show that the predicted gaze points can help to improve core vision tasks in FPV, including foreground object segmentation and action recognition. Our results suggest the importance of egocentric gaze for understanding objects and actions in FPV.

This work was in collaboration with Dr. Alireza Fathi and presented as an oral paper in ICCV 2013 [11]. This work had led to recent development on future gaze prediction in FPV [104] and inspired our subsequent work on FPV actions [12], which we will present in Chapter 5.

4.2 Overview

Our work is organized into two parts. The first part focuses on gaze estimation during object manipulation tasks. Our major contribution is leveraging the implicit cues that are provided by first person, such as hand location and pose, head/hand motion, for predicting gaze in FPV. We begin with an analysis of gaze tracking data from a wearable eye tracker and demonstrate that: (1) egocentric gaze is statistically different from on-screen eye-tracking; (2) there exists a strong coordination of eye, head and hand movements in the object manipulation tasks; (3) these coordinations can be used for predicting gaze in the egocentric setting. Moreover, we build a graphical model for gaze prediction that accounts for eye-hand and eye-head coordinations, and combines the temporal dynamics of gazes. The model requires no information of task or action, predicts gaze position at each frame and identifies moments of fixation. Our gaze prediction results outperform all previous bottom-up and top-down saliency detection algorithms by a large margin on publicly available datasets.

The second part demonstrates applications of predicted gaze in FPV. We provide extensive experimental results on two important applications in FPV: (1) foreground object segmentation and (2) egocentric action recognition. Simply by plugging in predicted gaze, we observe a significant performance boost in comparison to several previous methods. In object segmentation, the performance of our model is even comparable to alternative approaches that use ground-truth human gaze data.

4.3 Modeling Eye, Hand, Head Coordination

We focus on object manipulation tasks in a meal preparation setting, and explore the opportunity of gaze prediction using egocentric cues, including hand/head movement and hand location/pose. The coordination of eye, head and hand, as we show in this section, bridges the gap between these head and hand cues and gaze prediction.

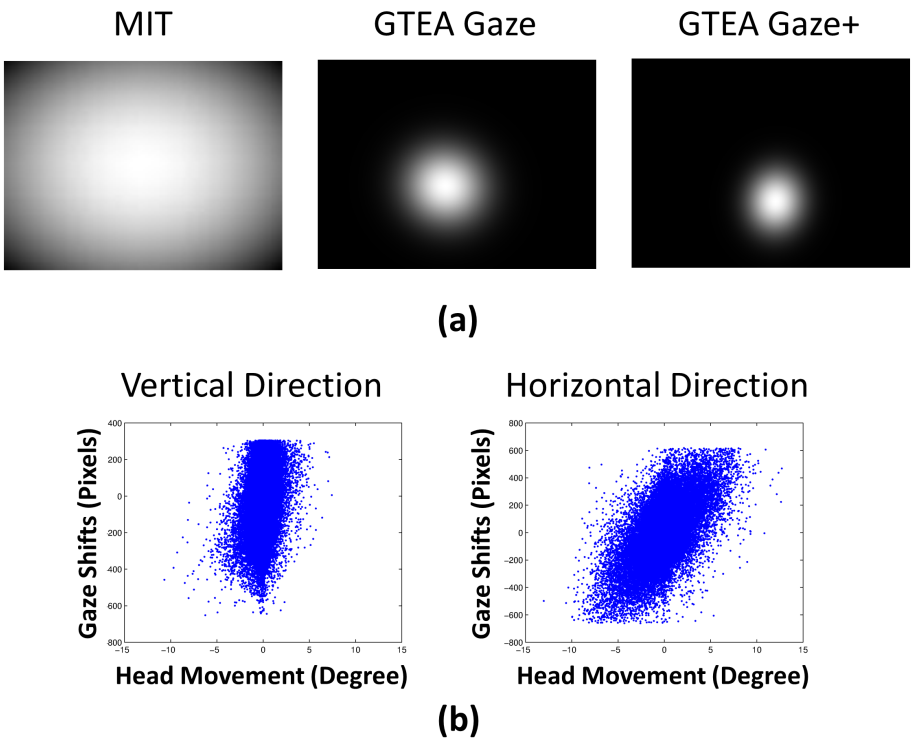

Figure 4.1: (a) Center bias (from left to right) for MIT saliency dataset, GTEA Gaze dataset and GTEA Gaze+ dataset. Egocentric gaze has a much smaller variance in space. Thus, head orientation provides a good approximation for gaze direction in egocentric videos. (b) A scatter plot of head movement against gaze shift along vertical and horizontal direction in GTEA Gaze+ dataset. The plot suggests a linear correlation in the horizontal direction.

Throughout the chapter, we use public dataset GTEA Gaze and a subset of GTEA Gaze+ (first 15 videos). While our Extended GTEA Gaze+ dataset is under construction, these are the only datasets that comes with gaze tracking dataset. Both datasets contain egocentric videos of meal preparation with gaze tracking and action annotations. We also consider MIT eye tracking dataset [125] for comparing gaze statistics. The MIT dataset includes gaze points from 15 subjects watching 1003 images on a screen.

Our modeling of eye, hand, head coordination is divided into two parts. First, we address the eye-head coordination by looking into gaze shifts and head motion. Second, we consider the eye-hand coordination by exploring the spatial relationship of gaze and egocentric hands.

4.3.1 Eye-Head Coordination

Several psychophysical experiments have indicated that eye gaze and head pose are coupled in various tasks [22, 128, 20, 129]. For example, large head movement is almost always accompanied by a large gaze shift. We explore the eye-head coordination in the object manipulation task by a data driven approach. The gaze statistics suggest a sharp center bias and a strong correlation between head motion and gaze shifts. These findings thus provide powerful cues for gaze prediction.

Egocentric Head Cues: We assume the camera is mounted on the first-person's head, continuously capturing the scene in front of the first-person. Instead of the absolute head orientation, we model the relative orientation of egocentric gaze with respect to the first-person's head. We estimate camera motion as a proxy to head motion using a 2D translational model. This is done by extracting 2D translation components in the homography, as we discussed in Section 3.3.1 in Chapter 3. The approximation, albeit less accurate, allows us to study the distribution of egocentric gaze in 2D.

Center Bias: Our first observation is a sharp center bias of egocentric gaze points [20]. We fit a 2D Gaussian as the center prior to all gaze points in GTEA Gaze and GTEA Gaze+ dataset, respectively, as shown in Figure 4.1. In comparison, we also visualize the center prior as a 2D Gaussian from MIT eye tracking dataset [125]. Egocentric gaze has a much smaller variance in the 2D image plane. This is due to the fact that egocentric vision captures a first-person's perspective in 3D world, where the gaze often aligns with the head orientation. In this case, the needs of large gaze shifts are usually compensated by head movements plus small gaze shifts. Thus, head orientation is a good approximation of gaze. Note that the preference of gaze towards the bottom part of the image is influenced by table-top object manipulation tasks.

Correlation between Gaze Shifts and Head Motion: We also observe a tight correlation between head motion and gaze shift in the horizontal direction. A scatter plot of gaze shifts (from the center) against head motion for GTEA Gaze+ dataset is shown in

Figure 4.1. The plot suggests a linear correlation in the horizontal direction, especially for large gaze shifts. Intuitively, a person tends to look at their right side if they turns their head towards right. This is again in consistent with the empirical finding in Figure 4.1. The correlation, therefore, allows us to predict gaze location from head motion.

4.3.2 Eye-Hand Coordination

Eye-hand coordination is an key component for performing object manipulation tasks. Eye gaze generally guides the movement of the hands to target [20]. Moreover, it has also been shown [130] that the proprioception of limbs may influence gaze shift, where the hands are used to guide eye movements. We use manipulation point to align gaze points with respect to the first person's hands, and discover clusters in the aligned gaze density map, suggesting a strong eye-hand coordination.

Encoding Egocentric Hands

To encode egocentric hands for gaze estimation, we explored two different ways of representing them. The first one explored the concept of a manipulation point. This representation abstracts away the pose, shape and location of hands using a single 2D point. The second one directly encode the confidence map using a histogram. This representation maintains a down-sampled version of the confidence map. Both representations is build upon hand segmentation results. Thus, we first describe our hand segmentation system, then present our representation of hands.

Manipulation Points: A major challenge for modeling first person's hand is how to represent hands with various poses. Instead of tracking the hand pose, we introduce manipulation point by analyzing hand shapes at each frame. A manipulation point is defined as a control point where the first person is mostly likely to manipulate an object using his hands. For example, for a single left hand, manipulation usually happens on right tip of the hand. For two intersecting hands, the manipulation point is generally around the intersecting part.

To find the manipulation point, we match the hand's boundary to configuration dependent templates. Examples can be found in Figure 4.2. A manipulation point provides an anchor with respect to current hand pose, and allows us to align image coordinate into the hand's coordinates. We use this representation for gaze estimation.

Finding manipulation points requires the distinction between left/right/intersecting hands. We build a simple classification model for doing this. or each hand region, we extract shape features (centroid, orientation of major axis, eccentricity and the area of the hand masks) and train a SVM to assign it to one of the three categories mentioned above. In addition, we assume there are at most two hands from the first person in a single frame. We greedily select at most two confident hand regions (with its area larger than a threshold). We also force mutual exclusiveness between region labels. For example, we can not assign a same label (single left/right hand/intersection hands) to more than one of the hand regions. And intersecting hands and single left/right hand can not show up simultaneously.

Histogram of Hands: The most straightforward way of representing hands is to use a down-sampled version of the hand confidence map, i.e. a histogram of hand confidence over regular 2D grids. We have later found it a fairly reliable representation if the hand map is accurate. In comparison to manipulation points, this representation maintain more information about the pose, shape and location of hands. We use this representation for gaze estimation and action recognition.

Gaze around Hands:

We segment the hand masks and extract the manipulation points of hands over each frame. We align the gaze points to the first-person's hands by setting the manipulation points as the origin (See Fig 3). The density maps of the aligned gaze points for four different hand configurations are plotted in Figure 4.3. For both GTEA Gaze and GTEA Gaze+ datasets, we observe high density around the manipulation point. The visualization suggest interesting spatial relationship between manipulation points and gaze points. For single

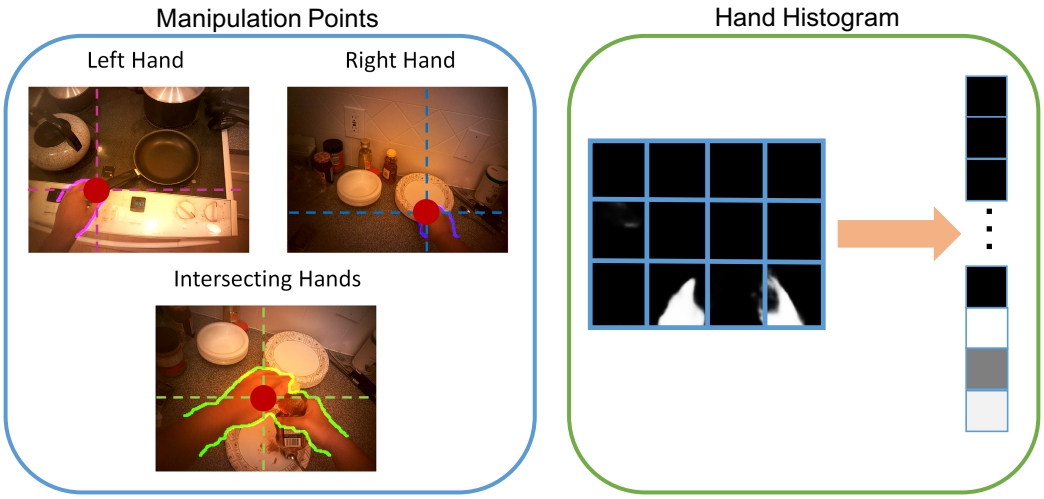

Figure 4.2: Two types of hand representations. Left: Hand segmentation and manipulation points (red dots). We classify hands into three different categories and show their correspondent manipulation points. The hands are colored by their configurations. Right: A simple histogram representation by laying out a regular grid over the image plane and taking the average scores within each grid. We found that it works well when hand segmentation results are accurate.

left/right hand, the gaze tends to fall on top right/top left region, where taking/putting actions might happen. For two separate hands, subjects are more likely to look in the middle, where the object usually stays. For two intersecting hands, gaze shifts towards the bottom, partly due to opening/closing actions. These spatial distributions are consistent with the observation that people tend to look at the object they are manipulating. Thus, they offer a simple cue for gaze prediction.

4.4 Gaze Estimation in Egocentric Video

We have demonstrate strong cues for gaze by the coordination of eye, hand and head movement. Now we present a learning based framework to incorporate all these egocentric cues for gaze prediction. The core of our method lies in a graphical model that combines egocentric cues at a single frame with a temporal model of gaze shifts.

Our gaze prediction consist of two parts: single frame gaze estimation and a temporal model of gaze. Our gaze estimation leverage on a novel set of first person visual cues that

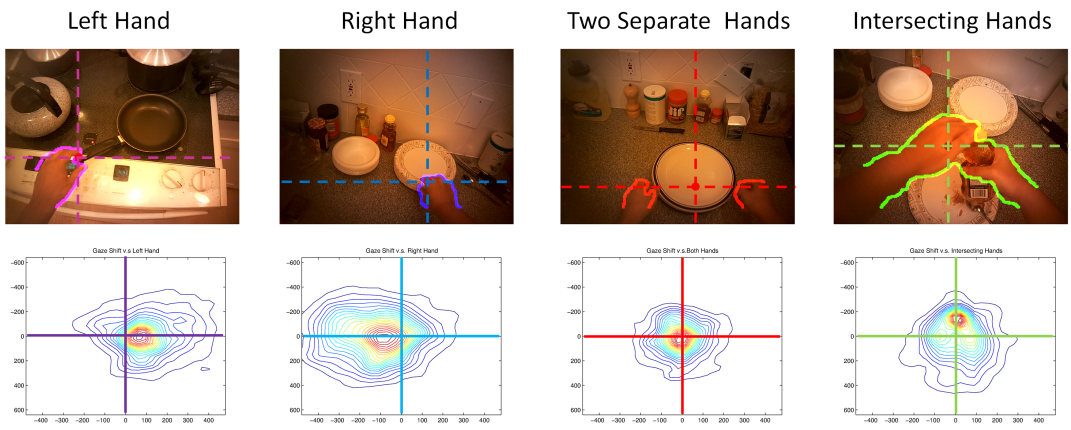

Figure 4.3: Top row: Hand segmentation and manipulation points (red dots). We present four different hand configurations and the correspondent manipulation points. The hands are colored by their configurations. Bottom row: Aligned gaze density map. We align the gaze points into the hand's coordinates by selecting the manipulation points as the origin, and projecting the gaze point into the new coordinate system every frame. We then plot the density map by averaging the aligned gaze points across all frames within the dataset. High density clusters can be found around the manipulation points, indicating spatial structures for eye-hand coordination.

encode head motion and hand locations. Our temporal model further infer identify fixations among all gaze points and refine single frame outputs. Specifically, fixation is defined as the pause of gaze within a spatially limited region ($0.5 \sim 1$ degree) for a minimum period of time ($80 \sim 120$ms) [131]. The modeling of fixations allow us to capture the temporal dynamics of gazes. We present our features and our model in this section.

4.4.1 Features

We extract egocentric features regarding the first person's hand and head cues. The feature vector z_t for frame t contains the manipulation point (2D), the global motion vector (2D), the hand motion vector (2D), the hand configuration (1D categorical). Therefore, for every frame, we get a 7 dimensional feature if hands are detected or a 3 dimensional feature if no hands are presented.

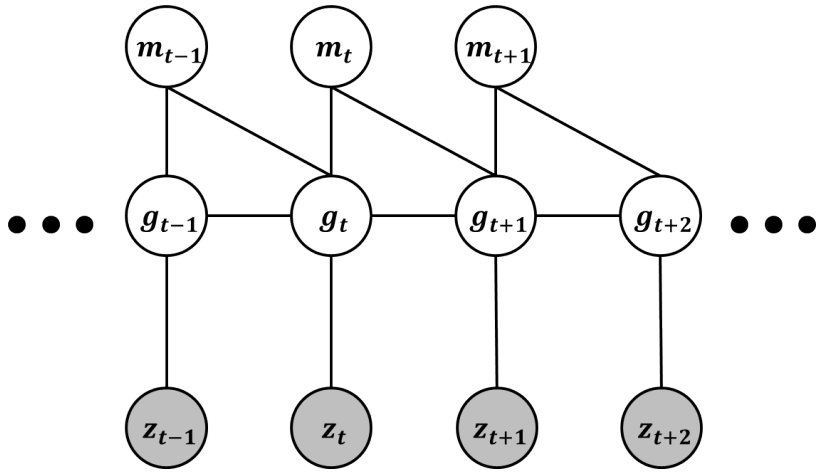

Figure 4.4: The graphical model of our gaze prediction method. Our model combines single frame egocentric cues with temporal dynamics of gazes. We extract features z_t at each frame t, predict its gaze position g_t and identify its moments of fixation m_t.

4.4.2 The Model

Denote the gaze point at frame t as $g_t = [g_t^x,\ g_t^y]^T \in R^2$ and its binary label as $m_t = \{0, 1\}$, where $m_t = 1$ denotes g_t is a fixation. Given egocentric cues $\{z_t\}$ for all frames $t = 1...K$, our goal is to infer the gaze points $\{g_t\}$ and its label $\{m_t = \{0, 1\}\}$. We model the conditional probability $P(\{g_t, m_t\}_{t=1}^K | \{z_t\}_{t=1}^K)$ as

$$P(\{g_t, m_t\}_{t=1}^K | \{z_t\}_{t=1}^K) = \prod_{t=1}^K P(g_t|z_t) \prod_{t=1}^K P(m_t|g_{N(t)}), \tag{4.1}$$

where $g_{N(t)}$ are the temporal neighbors of g_t. In our model, we set neighborhood to be two consecutive frames (133ms for GTEA Gaze and 80ms for GTEA Gaze+). The choice corresponds to the minimum duration of an eye fixation [20, 131]. The model consists of 1) $P(g_t|z_t)$ a single frame gaze prediction model given z_t; 2) $P(m_t|g_{N(t)})$ a temporal model that couples fixation m_t and gaze prediction $g_{N(t)}$. The graphical model is shown in Figure 4.4.

Single Frame Gaze Estimation: We use random regression forest for single frame gaze prediction. A random regression forest is an ensemble of decision trees. For each

branch node, a feature is selected from a random subset of all features and a decision boundary is set by minimizing the Minimum Square Error (MSE). The leaf nodes keep the mean value of all training samples that end up in the node. And the final result is the weighted average of all leaf nodes that a testing sample reaches. We choose random forest since our feature vector z_t might contain categorical data, which is easy to handle using a decision tree. We train two models for gaze prediction, one with only head cues and one with both hand and head cues. Our model will step back to using head motion cues if hands are not detected.

For simplicity, we train two regression forests for horizontal and vertical direction separately. The regression builds a map f between feature vector z_t to a $2D$ image coordinates $\tilde{g}_t = f(z_t)$, i.e. the prediction of gaze point at frame t. The probability $P(g_t|z_t)$ is then modeled as a Gaussian centered at $\tilde{g}_t$ with covariance $\Sigma_s \in R^{2\times 2}$

$$P(g_t|z_t) \propto \exp\left(-\|g_t - \tilde{g}_t\|^2_{\Sigma_s}\right), \tag{4.2}$$

where $\|g_t - \tilde{g}_t\|^2_{\Sigma_s} = (g_t - \tilde{g}_t)^T \Sigma_s^{-1} (g_t - \tilde{g}_t)$ is the Mahalanobis distance.

Fixations and Gazes: Gaze prediction and fixation detection are tightly coupled. On one hand, fixation m_t can be detected given all gaze points. On the other hand, there is a strong constraint over gaze locations if we know current gaze point is a fixation. For example, g_t should be close to g_{t-1} if $m_t = 1$. Therefore, we model the conditional probability $P(m_t|g_{N(t)})$ as

$$P(m_t|g_{N(t)}) \propto \exp\left(-m_t \sum_{i\in N(t)} \|g_i - g_t\|_2^2\right) \tag{4.3}$$

where m_i can be obtained by a fixation detection algorithm given gaze points $g_{N(t)}$. Here we use a velocity-threshold based fixation detection [131]: a fixation is detected if velocity of gaze points are below a threshold c over a minimum amount of time (two frames in our

case).

$$m_t = \prod_{i \in N(t)} \frac{-sign(\|g_i - g_t\|_2^2 - c) + 1}{2}, \tag{4.4}$$

where $sign(x) = -1$ if $x < 0$ and $sign(x) = 1$ if $x >= 0$.

4.4.3 Inference and Learning

Inference: To get the gaze points $\{g_t\}_{t=1}^K$ and fixations $\{m_t\}_{t=1}^K$, we apply Maximum Likelihood (ML) estimation of Eq (4.1). The minimization of negative log likelihood function is given by

$$\begin{aligned} \min_{\{g_t, m_t\}_{t=1}^K} & -\log(P(\{g_t\}_{t=1}^K, \{m_t\}_{t=1}^K | \{z_t\}_{t=1}^K)) \\ & = -\log\left(\prod_{t=1}^K P(g_t|z_t) \prod_{t=1}^K P(m_t|g_{N(t)})\right) \\ & = \sum_{t=1}^K \|g_t - \tilde{g}_t\|_{\Sigma_s}^2 + \lambda \sum_{t=1}^{K-1} m_t \|g_{t+1} - g_t\|_2^2 \\ s.t. \quad & m_t = \frac{-sign(\|g_{t+1} - g_t\|_2^2 - c) + 1}{2} \quad \forall t \end{aligned} \tag{4.5}$$

Projected gradient descent is used to obtain a local minimum of Eq (4.5). We first perform gradient descent over the object function assuming m_t is known and ignore the constraints. m_t is then updated to make all constraints feasible. These two steps run iteratively until convergence. Intuitively, the optimization follows a EM like updating by (1) identifying fixations m_t by velocity-thresholds given all gaze predictions g_t and (2) smoothing the gaze points g_t given fixation labels m_t.

Updating m_t given g_t is straightforward, we estimate m_t using Eq (4.4). Updating g_t given m_t is more challenging, since g_t and g_{t+1} are coupled together with m_t. Given m_t, we can rewrite Eq (4.5) using its matrix form. Let $G = [g_1 \ ... \ g_K]^T$, $\tilde{G} = [\tilde{g}_1 \ ... \ \tilde{g}_K]^T$ and $m = [m_1 \ ... \ m_K]^T$. Also we denote matrix A as the Toeplitz matrix correspondent to the

convolution kernels $[-1\ 1]^T$. The updating of G is equal to

$$\min_G \|G - \tilde{G}\|^2_{\Sigma_s} + \lambda \|m^T A G\|^2_2 \tag{4.6}$$

The solution is given by setting the first order derivative to zero

$$G^* = \left(\Sigma_s + \lambda A^T m m^T A\right)^{-1} \Sigma_s \tilde{G}. \tag{4.7}$$

Learning: Learning the model is relatively easy. We first train the single frame random regression tree, using 40 trees. The parameters needed to be determined now are the velocity threshold c, the covariance matrix Σ_s and the constant λ. We select c to be roughly the distance of 1 degree of angular error (50/80 pixels for GTEA Gaze and GTEA Gaze+ respectively). Σ_s defines the Mahalanobis distance between gaze points, and is learned by re-sending training samples into random forest and re-estimating the error covariance. We empirically select $\lambda = 0.4$.

4.4.4 Benchmark

We consider gaze estimation as a binary classification problem. For each frame, the locations of gaze points are marked as positive and the rest are negative. We then threshold the predicted gaze map and match it to our binary labels. This allow us to use two standard, complementary measures to assess the performance of our gaze prediction method: Area Under (ROC) Curve (AUC) and Average Angular Error (AAE). AUC measures the consistency between a predicted saliency map and the ground truth gaze points in an image, and is widely used in the visual saliency literature. AAE measures the angular distance between the predicted gaze point (e.g. the most salient point) and the ground-truth gaze, and is widely used in the gaze tracking literature. Since our method outputs a single predicted gaze point, we generate a saliency map that can be used for AUC scoring by convolving an isotropic Gaussian over the predicted gaze.

Egocentric gaze has strong center bias. Such a bias will affect the performance of gaze prediction, as also pointed out by Tatler et al. [132] and Zhang et al. [133]. To remove the bias, we use the AUC score by Zhang et al. [133]. For one image, the positive sample set is composed of gaze points on that image, whereas the negative sample set is composed of the union of all gaze points across all images and all subjects from the same data set, except for the positive samples. Each saliency map generated by different algorithms is thresholded and then considered as a binary classification problem that separate the positive samples from negative samples. At a particular threshold, we compare the binary saliency map with the positive and negative sample set to calculate the true positive/false positive rate. Sweeping over thresholds yields an ROC curve. AUC is the area under the ROC curve and provides a good measure of the power of the saliency map for gaze prediction. Chance level is 0.5, and perfect prediction is 1.0. Please refer to [133] for more discussions of different AUC scores.

4.4.5 Results

Both GTEA Gaze and GTEA Gaze+ dataset contain gaze data from eye tracking glasses, which are used as ground truth for gaze prediction. We compare our results with five competing methods: a baseline center prior prediction using 2D Gaussian, three bottom-up saliency detection algorithms (Itti and Koch [126], GBVS [127], Hou et al. [134]) and one top-down saliency algorithm [10]. For all the previous methods, we use the authors' own implementations for benchmarking purposes. The motion cues in [126, 127] are enabled for fair comparison. One issue is that our previous method [10] requires action labels for gaze prediction. We supply their method with ground truth action labels in all of our experiments. We emphasis that our method uses neither bottom-up features nor top-down action labels.

For GTEA Gaze dataset, we use the same training (13 videos) and testing (4 videos) split as [10] for fair comparison. For GTEA Gaze+ dataset, we perform a five-fold cross

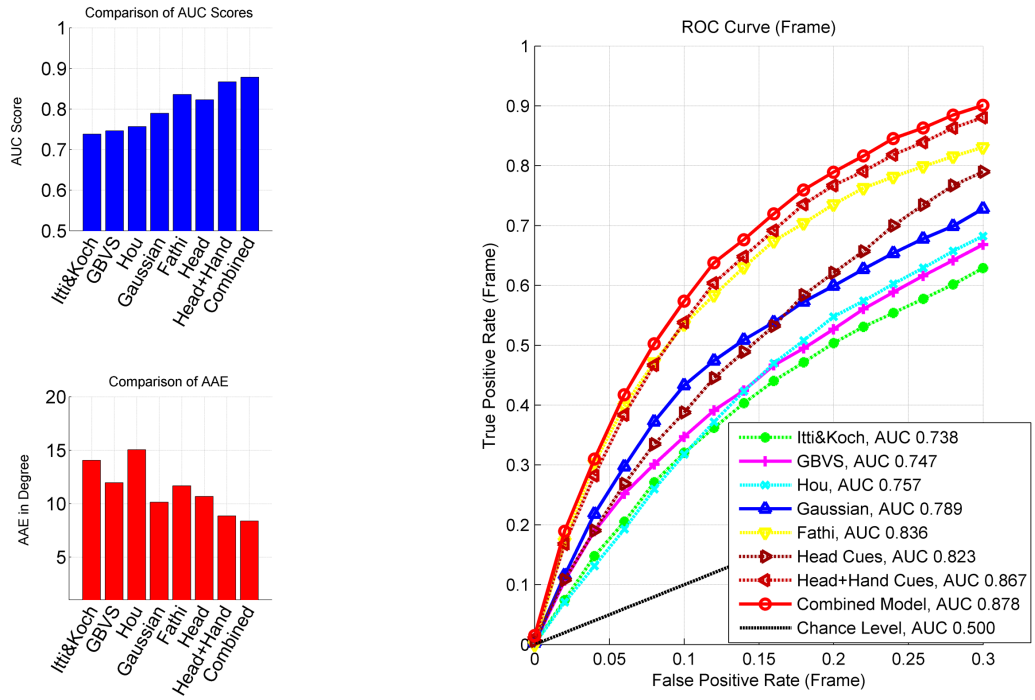

Figure 4.5: Left: AUC scores and AAE for 8 different methods in **GTEA Gaze dataset**. Our combined model achieves the highest AUC score (87.8%) and lowest AAE (8.35 degrees) among all methods. Our method consistent generates more accurate predictions. We has less AAE than [10] for 75% of all frames (67% for 2D Gaussian). Right: ROC curve for different methods. Our method requires no information about action or task, and largely outperforms the bottom-up and top-down gaze prediction method.

validation by using 4 subjects for training and 1 subject for testing. For all our results, we average over 10 runs of random forest.

Figure 4.5 shows the quantitative comparison of AUC, AAE and the ROC curve in GTEA Gaze. Overall, our combined model achieved AUC score of 87.8%, where the state-of-the-art [10] gives 83.6% by using the ground truth action labels in testing. Our method also ranks highest for AAE with 8.35 degrees, where the second best is 2D Gaussian (10.16 degrees). Our method works surprisingly well and outperform the sophisticated top-down method [10] by 4.2%. Our method with head cues achieved AUC score of 82.3% and AAE of 10.68 degrees. Adding hand cues significantly improved the score (86.7%) and reduce the angular error (8.85 degrees). Our temporal model added another 1% of AUC and 0.5 degree of AAE.

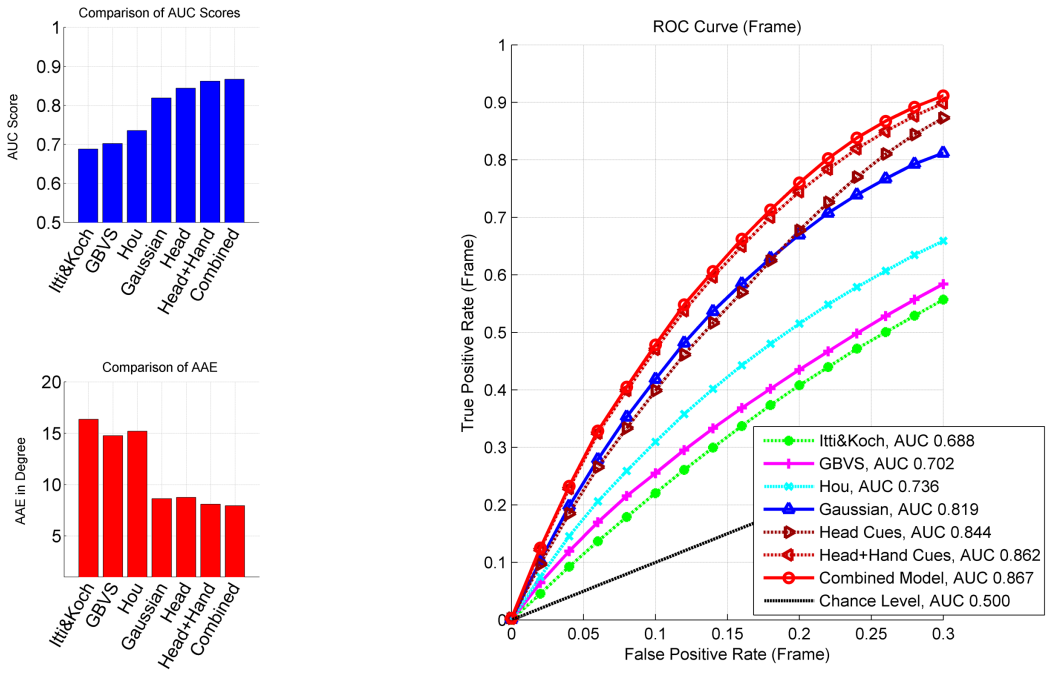

Figure 4.6: Left: AUC scores and AAE for 7 different methods in **GTEA Gaze+** dataset. Again, our methods outperform all other methods in both AUC score and AAE. Our method has less AAE than second best (2D Gaussian) for 69% of all frames. Right: ROC curves. It is interesting to find that the 2D Gaussian consistently outperform bottom-up methods.

Another interesting finding is that the center prior gives better accuracy than all of the bottom-up results in AUC and has a reasonable AAE. These results suggest that egocentric cues can provide a reliable gaze estimate without high-level task constraints and low-level image features. Our method benefits from using the strong egocentric cues (head, hand and eye coordination) for gaze prediction and bypasses the challenging object segmentation step required by [10]. We argue that modeling the coordination is simpler and more reliable than modeling of the high level tasks for gaze prediction.

We also tested our method in GTEA Gaze+ dataset. The results, including AUC, AAE and the ROC curve are shown in Fig 6. Again, our method has the best AUC of 86.7% and the best AAE of 7.93 degrees, outperforming the second best (2D Gaussian) by 4.8% and 0.7 degrees respectively. Using head motion already outperformed the center prior and adding hand cues further improves the results. Again, the center prior performs better than bottom-up methods. One possible explanation is that bottom-up saliency may be an effec-

tive predictor for visual search tasks, where image features may naturally draw the viewer's attention during a scan. But for hand-eye coordination tasks the gaze is coordinated with the head, making the head orientation a more effective approximation.

The GTEA Gaze+ dataset also provides ground truth labels for fixations, which are produced by the eye tracking glasses. We evaluated our method for fixation detection, and found that it achieved 84.7% accuracy.

Analysis of Egocentric Hands

In our previous work, we represented egocentric hands using manipulation points based on hand masks from TextonBoost [119]. This hand crafted representation is rather ad-hoc yet generates good results for gaze prediction. We had explored using a more principled hand histogram based on TextonBoost masks, yet the results are less satisfactory. After the publication of the work, we have replace TextonBoost with our new hand segmentation pipeline using deep models (see Section 3.3.2 in Chapter 3), which produces much more accurate hand segmentations.

Accurate hand masks allow us to explore different ways for representing hands, and further improved the results on gaze estimation. We made the following improvement. First, we use FlowNetV2 [118] to get a more reliable estimation of 2D head motion. Second, we replace the manipulation points with a hand histogram of the size 4×3. For a fair comparison, we train and test our model on the same video clips. Our improved method achieved AUC of 88.2% and AAE of 7.52 degrees. The improvements are 1.5% of AUC and 0.39 degrees of AAE. While the improvement is not significant, this is an important step. These results suggest that we can get rid of the heuristic representations, and still keep improving the accuracy when the underlie signals (e.g. hand masks) are accurate. A further step is to replace random forest with deep models, such that we can directly learn from raw inputs. We leave this for future work.

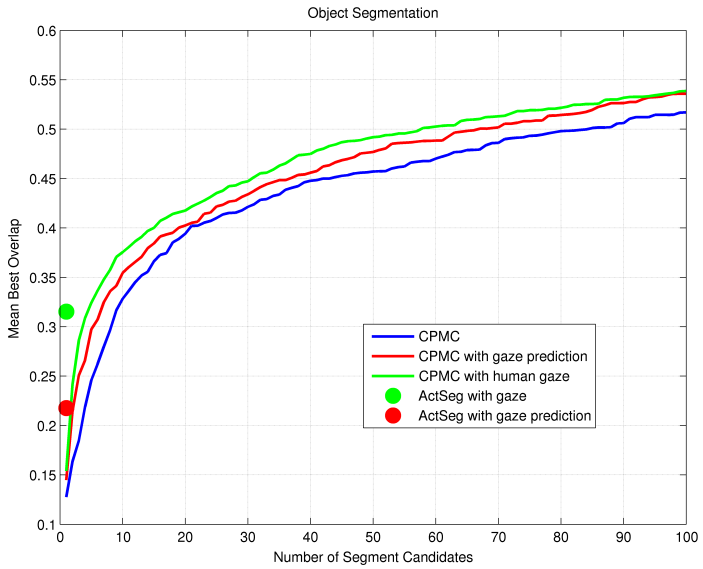

Figure 4.7: Foreground object segmentation results. We plug in our gaze prediction into two different algorithms. For ActSeg, human gaze achieves 31.5% and our gaze prediction reaches 21.7%. CPMC achieves the same score by the first 4 segments with the help of human gaze, and by the first 6 segments using our gaze prediction. We also improve CPMC results by 2.6% over top 100 segments using gaze, with only a small performance gap between human gaze and our predicted gaze.

4.5 From Gaze to Objects and Actions

4.5.1 Object Segmentation

We further demonstrate that gaze prediction can be used to segment task-relevant foreground objects. In a task-oriented setting, we define the foreground object as the one that is involved in current task. We supplement a subset of action clips with object masks, and study how egocentric gaze can help to find foreground objects.

Annotation

We design a protocol to obtain the ground truth annotation. First, the video is cropped into small clips that last for 1.5 second. We pick video clips within actions that involve objects (e.g. take/put, open/close) and show each clip to an annotator for segmenting the foreground object. And we obtained 234 object masks from 300 video clips selected from

6 of the videos in GTEA Gaze+.

Our annotator is asked first "Is there an object involved in the video clip given current task?". If yes, the annotator is required to pick a frame within the short clip and segment the object. Note the annotator can choose to skip one clip if no object exists, or segment more than one object within a single clip.

We recruit two annotators and ask them to annotate a small subset (50 video clips) independently. The annotators report a high confidence for most of the videos and their annotation consistency is high. For 42 out of 50 videos, they specify the same object. Then we let them annotate the rest of the dataset. We obtain 234 object masks in different frames with correspondent gaze points from over 300 video clips across 6 videos in GTEA+. Examples of annotations can be found in Figure 4.8. Our goal is to segment the foreground object given a single frame and gaze information, either by eye tracking or gaze prediction.

Egocentric Objects and Egocentric Gaze

We studied the relationship between foreground object masks and fixation points in our dataset. We found that 82.9% (194/224) of our object annotations contain a fixation in the same frame. And 75.2% of the fixations lies within the foreground object boundary. Moreover, 94.3% of the fixations lies within the 80 pixels (1 degree) from the nearest foreground object boundary. The statistics suggest that human gaze tends to focus on task-relevant objects [20]. However, it is not always true that the fixation lies in the object boundary [124]. Possible explanation includes micro saccade [20] or gaze tracking error.

Segmenting Objects using Gaze

We used both our gaze prediction method and the ground truth gaze point to seed two different methods for extracting foreground object regions: ActSeg [124] and CPMC [135]. Given the ground truth segmentation, we score the effectiveness of object segmentation under both predicted and measured gaze, thereby obtaining an alternate characterization

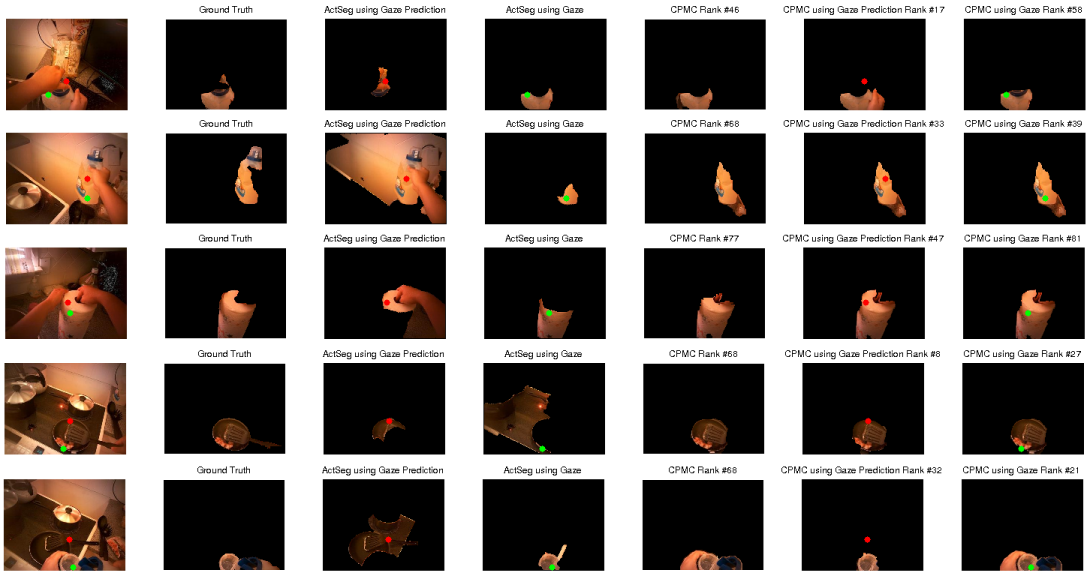

Figure 4.8: Examples for object segmentation results. Green dot: human gaze; Red dot: our predicted gaze. From left to right: the original image; the object annotation; ActSeg result using predicted gaze; ActSeg result using human gaze; best CPMC result within first 100 segments; best CPMC result within first 100 segments using predicted gaze; best CPMC result within first 100 segments using human gaze. ActSeg achieves 31.5% and 21.7% overlapping score for human gaze and predicted gaze, respectively. CPMC get equivalent performance to ActSeg from the first 10 segments. We improve CPMC results by 3% for the first 100 segments using gaze.

of the effectiveness of our gaze prediction method. ActSeg [124] takes gaze points as the input and outputs one segment per gaze point. It assumes that the gaze point always lies within the object boundary and segments an object by finding the most salient boundary. CPMC [135] uniformly samples seed points across the entire image, then generates object hypothesis and ranks them. We modified the implementation of CPMC to put dense seeds only in the vicinity of the gaze point.

We score result segments by the mean best overlapping scores, defined as the average of best overlap (interaction over union) between a segment and the ground truth. We measure the performance of CPMC by selecting top K candidates and varying the number of K. The results are reported in Figure 4.7. For ActSeg, human gaze gives 31.5% and our gaze prediction gives 21.7%. For CPMC, we get equivalent performance to ActSeg from the first 6 segments, and then improve the results by 2.6% by using gaze with the first 100

segments. The performance using our gaze prediction method is comparable to that using ground truth gaze.

Moreover, we show qualitative comparisons in Figure 4.8. ActSeg assumes the gaze point lies in the object and outputs one segment per gaze point. CPMC generates object hypothesis based on foreground seeds and rank them afterwards. The assumption of ActSeg is not always true (See 4th row for a counter example) yet the method produces reasonable results. The best CPMC segment is generally better than ActSeg. We improve the ranking of the best object segment in CPMC using gaze information for most of the cases. The performance using our predicted gaze is comparable to that using ground truth gaze.

4.5.2 Action Recognition

Egocentric gaze is not only useful for foreground object segmentation, but also helps to recognize first-person's action. We show that action recognition accuracy can be significantly improved by plugging our predicted gaze into our previous work [10].

Action Recognition using Gaze

To make the section self-contained, we briefly describe our previous method here. For more details, we refer to our paper [10]. The key idea of the method is that an action can be inferred from the local image features observed in the vicinity of the sequence of fixation points. Our feature set includes (1) appearance features of color and texture; (2) object features from semantic segmentation [119]; (3) future manipulation features that back-propagate future foreground regions into the current frames. These image features are extract at each frame. A classifier is trained to map the features to action categories, followed by a Hidden Markov Model that smooth the temporal output and produce the final action labels.

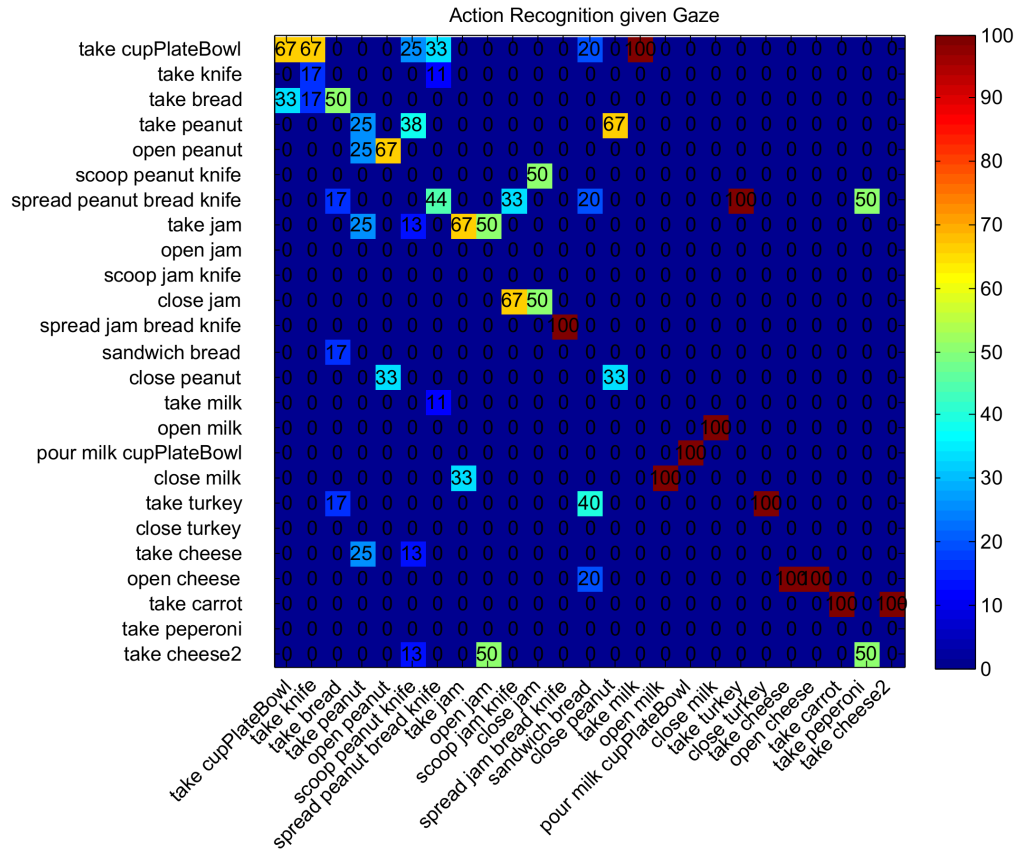

Figure 4.9: Confusion matrix of action recognition using predicted gaze on GTEA Gaze dataset for 25 classes. The average accuracy is 32.8% in comparison to 29% in the previous method.

Results

Our action recognition results are summarized by the confusion matrix in Figure 4.9. We compare our results against [10]. Using our gaze prediction, we improve the action recognition result to 32.8% from the state-of-the-art [10] at 29%. The upper bound of the method is given by human gaze at an accuracy of 47%. For 7 out of 25 classes, we perform better than [10]. Again, we can not report results on GTEA Gaze+ due to the lack of object annotations. We notice the large gap between our gaze prediction and real human gaze. We conclude the gap is largely due to the fact that our previous method [10] is sensitive to input gaze. We have further developed methods on using gaze for FPV action recognition [12] and will present them in next chapter.

4.6 Conclusion

In this chapter, I present a novel model for estimating FPV gaze by modeling the coordination of head, hand and gaze in egocentric actions. Our results show for the first time that it is possible to estimate egocentric gaze, by using an "embodied" representation of first person visual cues and without the need for object or action information. More specifically, our method estimate FPV using head motion and hand locations. I further propose a sequential model to capture the dynamic behavior of FPV gaze. Our gaze prediction results outperform previous methods by a large margin on GTEA Gaze and GTEA Gaze+ datasets.

Moreover, I demonstrated that egocentric gaze can be used to understand objects and actions in FPV. I show a significant performance boost in recognizing actions and segmenting foreground objects by plugging in our predicted gaze into existing methods. Our results suggest that egocentric gaze index critical regions of objects and actions in FPV.

It is important to note that our method is limited to object manipulation tasks, where hands are often visible. An interesting opportunity and an immediate next step is to explore gaze estimation in other daily activities. For example, how can we estimate gaze when the person is driving or walking or playing soccer? Task specific information, such as objects, actions or the body pose of the person, might be required to understand the gaze behavior in these settings.

CHAPTER 5
FIRST-PERSON ACTION RECOGNITION

Understanding human actions from videos has been a well-studied topic in computer vision. The recent advent of wearable devices has led to a growing interest in understanding egocentric actions, i.e. analyzing the first person's behavior using egocentric videos. Since an egocentric camera is aligned with the wearer's field of view, it is primed to capture the first person's daily activities without the need of instrument the environment. Knowledge of these activities facilitates a wide range of applications, including remote assistance, mobile health and human-robot interaction.

Despite tremendous efforts on understanding actions in surveillance settings [26, 27], it remains unclear if previous methods of action recognition can be successfully applied to first person videos. Our observation is that first person video includes frequent ego-motion due to body movement. This camera motion can potentially hamper the motion-based representations that underlie many successful action recognition systems. In contrast, state-of-the-art egocentric action recognition methods [75, 51, 10] rely mainly on an object-centric representation for discriminating action categories. However, most of their works did not test motion-based representations on a common ground, e.g. separating the foreground motion from the camera motion. Thus, a systematic evaluation of motion cues in egocentric action recognition remains missing.

What makes egocentric videos different from surveillance videos? The key is not simply that a camera is moving, but rather a person who is wearing the camera. In a natural setting, the camera wearer performs an action by coordinating his body movement during an interaction with the physical world. The action captured by an egocentric video contains a rich set of signals, including the first person's head/hand movement, hand pose and even gaze information. We consider these signals regarding the first person as mid-level

egocentric cues. They usually come from low-level appearance or motion cues, e.g. hand segmentation or motion estimation, and are complementary to traditional visual features. These mid-level egocentric cues reveal the underlying actions of the first person, yet have been largely ignored by previous methods of egocentric action recognition.

We provide an extensive evaluation of motion, object and egocentric features for egocentric action recognition. We set up baselines using two different video representations: the traditional local descriptors from Dense Trajectories (DT) [35], the more recent deep convolutional networks from Temporal Segment Networks (TSN) [39]—a variant of Two Stream Networks [7]. These are successful video representations for action recognition in a surveillance setting. We then systematically vary the methods by adding motion compensation, object features and egocentric features. Our benchmark demonstrates how these choices contribute to the final performance. We identify a key set of practices that produce statistically significant improvement over the state-of-the-art methods. In particular, we find that simply extracting features around the first-person's attention point works surprisingly well. Our findings lead to a significant performance boost over state-of-the-art methods on major datasets.

This chapter is organized as follows. In Section 5.2, we summarize our contributions. In Section 5.3 and 5.4, we present our methods using DT and deep models, respectively. Finally, we conclude the chapter in Section 5.5.

5.1 Contributions

This chapter has three major contributions.

- I demonstrate that how egocentric cues can be combined with low-level features or deep features to effectively improve the performance of FPV action recognition.
- I established the first systematic evaluation of motion, object and egocentric features for FPV actions. Our benchmark uses two different types of video representations,

and uncovers the importance of egocentric cues in FPV action recognition.

- My study identifies a key set of ingredients that are critical to the performance. Our work not only offers the best practice with significant performance boosts over major datasets, but also leads to valuable insights for understanding FPV actions.

This work was a collaboration with Dr. Alireza Fathi, Zhefan Ye and Yun Zhang. The work was published in ECCV 2012 [10] and CVPR 2015 [12]. Our key idea of using egocentric cues for FPV actions is further developed in several recent vision papers [79, 80, 81].

5.2 FPV Action Recognition using Dense Trajectory

In this section, we present the details of our action recognition pipeline build on top of DT. Figure 5.1 provides an overview of our approach. We will start by introducing motion, object and egocentric cues, followed by our pipeline for recognition, and finally present our results and findings.

5.2.1 Motion, Object and Egocentric Cues

We give a brief description of the motion, object and egocentric features used in this chapter. Different encoding schemes for the egocentric features are also discussed.

Local Descriptors for Motion and Object Cues

Our method is built upon the pipeline of DT [35], which has not been fully explored in egocentric video. The success of DT lies in its dense tracking strategy using optical flow, and the combination of multiple descriptors aligned with the trajectories. Dense sampling ensures that key visual information is captured by the trajectories. The feature set is designed for different aspects of an action, including trajectory shape, 2D image boundary, motion direction and motion boundary. Each descriptor in DT thus includes its spatial-temporal

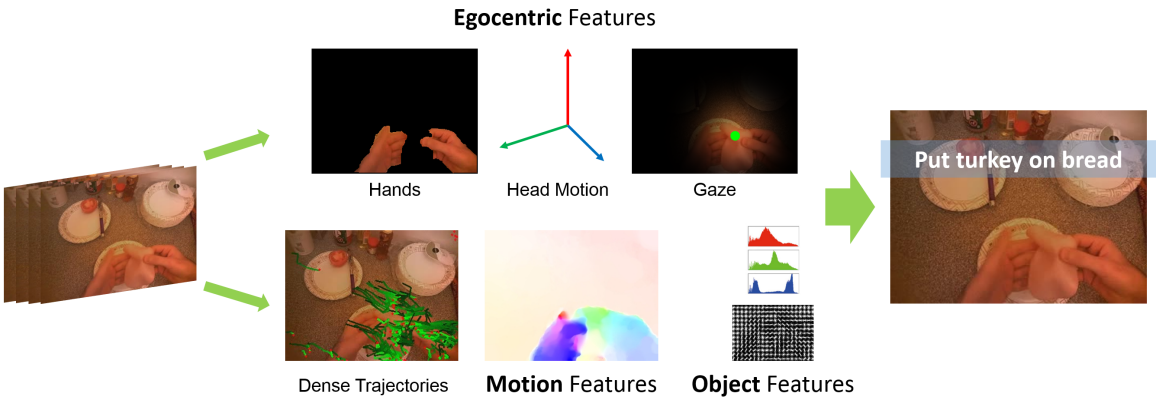

Figure 5.1: FPV action recognition pipeline using Dense Trajectory. We propose to combine a novel set of first person visual cues with low-level object and motion cues for recognizing egocentric actions. Our *egocentric* features encode hand pose, head motion and gaze direction. Our *motion* and *object* features come from local descriptors in Dense Trajectories, with proper motion compensation. We design a systematic benchmark to evaluate how different types of features contribute to the final performance, and seek the best recipe using motion, object and egocentric cues. Our findings significantly advance the results in major benchmarks.

trajectory and the **features** along the trajectory. We separate these features into motion features and object features.

Motion Features: Motion is the inherent nature of an action. DT captures motion information by 1) trajectory features of the shape of a trajectory; 2) Histogram of Flow (HoF) as the local motion pattern; 3) Motion Boundary Histogram (MBH) using the gradient of optical flow split into vertical (MBHy) and horizontal directions (MBHx), as the shape of moving foreground objects.

Object Features: Object information is crucial in egocentric settings, as many of our actions involve the interaction with objects. DT include Histogram of Oriented Gradient (HOG), which encodes the 2D image boundaries. We further augment DT with histogram of LAB color and Local Binary Patterns (LBP) along the trajectory, capturing color and texture information. We deliberately choose our object features as low-level descriptors to keep our pipeline simple.

Egocentric Cues for Actions

We introduce our egocentric cues, which are used as mid-level features, and show how they can be inferred from an egocentric video.

Hand, Head and Gaze: As discussed in Section 3 of Chapter 3, we extract a set of first person visual cues over each frame, including head motion (homography), hand manipulation points and 2D gaze points given by eye tracker.

Encoding Head and Hand Movement: Head motion and hand movement is complementary to visual features. We thus directly encode the head motion and the trajectory of manipulation points as separate feature channels. We also experimented with encoding the trajectory of gaze points, yet only found negligible improvement.

Egocentric Cues Meet Local Descriptors

The key challenge for using local descriptors in egocentric video is that they often fire at locations that are irrelevant to the current action. This is mainly due to camera motion and background clutter. In addition to encode egocentric cues independently, we show that they can be used to produce meaningful local descriptors for egocentric actions. This is done by motion compensation and trajectory selection.

Motion Compensation: Motion compensation is important for egocentric actions. Several recent efforts addressed the issue in a surveillance setting by either stabilizing the input video [136] or compensating the optical flow [137, 138]. The latter has been proved effective for action recognition [137, 138]. Thus, we adapt a similar technique in [137]. We back-warp a future frame using the homography, re-estimate optical flow for motion features, and reject trajectories with small motion.

Motion compensation has two major effects. First, it helps to select trajectories on foreground regions that move differently from the camera motion. Secondly, it helps to generate more reliable motion features that exclude the ego-motion from the dense optical flow filed. Note that our implementation of motion compensation is different from [137].

Our version uses ORB features homogeneous distributed on the image plane for estimating homography, only need to compute dense optical flow once, and is thus more efficient with comparable results.

Trajectory Selection: Gaze points index key locations that are discriminative for actions. We [10] have previous proposed a simple heuristic by only encoding visual features around gaze points. In this case, egocentric features provide a weak spatial prior of an action. We experiment with selecting local descriptors by its trajectories in the vicinity of both manipulation point and gaze point. Trajectory selection drives local descriptors to focus on egocentric actions, by filtering out descriptors with irrelevant trajectories, e.g. trajectories lies on the background clutter. It also improves the efficiency as less descriptors are used for recognition.

5.2.2 FPV Action Recognition Pipeline

We now describe our approach to combine object, motion and egocentric features for action recognition. We discuss the details of our implementation and benchmark, followed by our results and findings. We achieve a significant performance boost on major datasets.

Method and Implementation

Feature Extraction: Our method shares a similar pipeline with [137]. We track feature points using DT in an input video, using a time window of 6 frames. Note the trajectory length is shorter than [137], as many of the egocentric actions only last for a few seconds. We extract a set of local descriptors aggregated along the trajectories. Each descriptor consists of 7 feature channels, including trajectory features, MBHx, MBHy, HoF, HoG, LAB color histogram and LBP. We use 8-neighbor comparison for LBP and quantize three color channels (LAB) separately into 8 bins each. Each trajectory is further divided by $2 \times 2 \times 3$ grids and histograms of features within each grid are concatenated. The final dimensions of the descriptors are 12 for Trajectory, 96 for HoG and LBP, 108 for HoF, 192

for MBH and 288 for Color. Other parameters of DT are kept the same as in [35]. We also extract egocentric features at each frame, including head motion parameters (8D) and hand manipulation point (2D).

Fisher Encoding: We encode all descriptors using Improved Fisher Vector (IFV) and concatenate the result vectors. IFV [139] has been shown to outperform other encoding methods in action recognition [140]. IFV is obtained by soft quantization of the projected descriptor of dimension D using a Gaussian Mixture Model (GMM) of K components. Zero, first and second order differences between each descriptor and its Gaussian cluster mean are calculated, and weighted properly by the Gaussian soft-assignments and covariance. They are then averaged into an unnormalized fisher vector. We further take a signed square root of its scalar components and normalize the vector with a unit l_2 norm [139]. The result is a fisher vector of the dimension $(2D + 1)K$. For all of our experiments, we first perform PCA to reduce the input feature dimensions by 50%, followed by a GMM with $K = 50$ using 200K randomly sampled descriptors. To eliminate randomness in clustering, all results are obtained by averaging over 5 runs.

Classification: We concatenate the IFVs from different features into the final representation of the video. We train a linear SVM over the final FV for action recognition. The SVM parameter C is selected by leave-one-subject-out cross-validation on the training set on GTEA (best C=40) and GTEA Gaze+ (best C=60). We manually set $C = 60$ for GTEA Gaze, where cross-validation is not feasible. For all dataset, we use a class-weighted SVM [141], such that each sample is weighted based on the frequency of its label. This is equivalent to re-weighting the errors for each class.

Implementation Details: We also implement spatial FV (SFV) [142] and data augmentation [143]. For IFV, We randomly sampled 200K local descriptors for GMM clustering with 50 components. Data augmentation is done by mirroring the videos horizontally in both training and testing. Our final classification results are given by the averaged score between the original video and its mirrored version. We find that SFV and data augmentation

consistently improve the performance, and include them in all methods across experiments.

Dataset and Baselines

We use three datasets for our experimental evaluations: GTEA, GTEA Gaze and GTEA Gaze+. They are publicly available and include action annotations. Each action consist of a verb and a set of nouns, such as "put turkey (on) bread". We choose these datasets because 1) they are designed for egocentric action and activity recognition; 2) they are captured by head-mounted cameras of human subjects with a set of rich egocentric cues.

Results have been reported on GTEA and GTEA Gaze in [76, 10, 11, 75]. However, they are not built upon a fair ground and can not be directly compared in order to properly understand the performance. This is due to the facts that (1) the action annotation of GTEA does not include all actions that start with the verb "put", which biases the benchmark; (2) Results in [10, 11] are reported over a subset of all actions in GTEA Gaze, again, missing all "put" actions; (3) GTEA Gaze+ includes over 900 categories in total, yet most of them happen only 1-2 times and no previous result has been reported; (4) No cross-validation is performed for the reported results, with the danger of over-fitting.

We establish the first rigorous baseline and evaluation criteria for these datasets. This is done by (1) re-annotating GTEA dataset to include actions that include the verb "put"; (2) reporting leave-one-subject-out cross-validation results on both old and new list of categories on GTEA; (3) reporting benchmark results on both partial and full list of GTEA Gaze; (4) defining the list of action categories on GTEA Gaze+ by requiring an action happens at least twice for each subject, leading to 44 action classes with 1958 instances; (5) providing leave-one-subject-out cross-validation results on GTEA Gaze+; (6) comparing the results on three datasets with a large set of baselines on a common ground. We also supplement the datasets with 2.5k hand masks. These masks are used to train our hand segmentation pipeline. Action annotations together with hand masks are publicly available at our project website.

Table 5.1: Our results on first person action recognition, grouped into four parts, with all numbers in percentages. The first group (row) includes the baselines of STIP, Cuboids, DT and IDT. In the second group, we compare motion (M) and object (O) features. Note our motion features is a subset from IDT with trajectory features, HoF, MBHx and MBHy. The third part focuses on egocentric features. We consider direct encoding of egocentric cues (E), as well as feature extraction around an attention point given by hand (H) or gaze (G). In the fourth part, we explore the combination of motion (M) and object (O) features with the attention point by hand (H) or gaze (G). By systematically varying different components, we uncover ingredients for egocentric action recognition and significantly advance the state-of-the-art results. (*Results are obtained using human gaze)

	GTEA(61) FixSplit	GTEA(61) CrossVal	GTEA(71) CrossVal	GTEA Gaze(25) FixSplit	GTEA Gaze(40) FixSplit	GTEA Gaze+(44) CrossVal
STIP	32.9	31.1	25.3	26.3	23.8	14.9
Cuboids	11.2	12.5	13.3	20.1	20.6	22.7
DT	33.0	34.1	32.9	34.2	34.1	42.4
IDT	39.8	42.5	40.5	41.3	27.7	49.6
M	37.3	39.6	38.7	40.3	27.5	45.6
O	56.7	53.9	55.0	42.5	28.2	53.4
O+M	56.9	56.1	55.2	43.2	29.5	56.3
Ego Only (E)	15.3	16.3	16.5	19.9	17.4	22.3
O+M+E	59.4	55.9	55.7	44.5	32.0	56.7
O+M+E+H	61.1	59.1	59.2	53.2	35.7	**60.5**
O+M+E+G	N/A	N/A	N/A	**60.9***	**39.6***	**60.3***
M+E+H	40.8	43.1	42.3	47.6	30.3	53.2
O+E+H	**66.8**	**64.0**	**62.1**	51.1	35.1	57.4
M+E+G	N/A	N/A	N/A	44.1*	33.1*	51.3*
O+E+G	N/A	N/A	N/A	53.4*	34.1*	57.7*
State-of-the-art	39.7 [76]	N/A	N/A	32.8 [11] 47.0* [10]	N/A	N/A

Our baselines include STIP [34], Cuboids [38], DT and Improved DT (IDT) [35, 137]. Note that we supplement IDT with our head motion estimation, which provides slightly better results in egocentric videos. We also include results from [76, 10, 11]. Our results are obtained by adding motion compensation (IDT [137]), object features and egocentric features on top of DT. We report average class accuracy as the benchmark criterion. For efficiency, we resize the videos into 320×240 for GTEA Gaze and GTEA Gaze+ dataset. We use the rectified frames for GTEA from [75] and resize the video to 360×203. We also reduce the frame rate by half for all datasets. Further increasing resolution or frame rate has negligible differences in results.

Results and Findings

For a fair comparison to previous work, we provide benchmark results in 5 different settings: (1) GTEA dataset with old labels using the same training and testing split as in [76, 75]; (2) GTEA dataset with old labels using leave-one-subject-out cross-validation; (2) GTEA dataset with new labels and leave-one-subject-out cross-validation; (3) GTEA Gaze dataset with the same action categories and training testing split in [10, 11]; (4) GTEA Gaze dataset with all action categories using the same training testing split in [10, 11]; (5) GTEA Gaze+ dataset with leave-one-subject-out cross-validation.

In particular, we divide the features into three parts and benchmark them separately: (1) **M**otion features obtained by concatenating FVs from trajectory features, MBHx, MBHy and HoF; (2) **O**bject features by concatenating FVs from HoG, LAB color histogram and LBP; (3) **E**gocentric features by concatenating FVs from head motion and manipulation point. We also denote H and G as selecting local descriptors using manipulation point and gaze point. Our results are summarized in Table 5.1. Best results are highlighted.

Imbalanced Data: We notice that both GTEA and GTEA Gaze have very few number of instances ($3 \sim 4$) for many categories. More precisely, the distribution of instances within each category is highly imbalanced. For example, in GTEA, while the action of "take bread" has 28 instances, 33 out of the 71 categories have less than 5 instances. This can produce misleading results [144], as missing one instance in these "sparse" categories can impose a large penalty of average class accuracy. There is no good way to fully resolve this issue. We use a class-weighted SVM to reweight the error for each class, which we found work well for our pipeline. Another possibility is to resample data points at the beginning using either under-sampling or over-sampling.

The same heavy-tailed distribution also holds for GTEA Gaze+, yet in a much better condition. With more instances, GTEA Gaze+ has a median number of 25 instances per category in comparison to 8 (GTEA) and 5 (GTEA Gaze). Therefore, it is less likely to get penalized in average accuracy by missing a few instances. Moreover, GTEA Gaze+ is

collected in a real kitchen setting with a higher resolution, while both GTEA and GTEA Gaze are captured at a lab environment. In this section, we report results on all datasets. However, we highly recommend future benchmarks on GTEA Gaze+, leaving GTEA and GTEA Gaze as proof of concept.

Motion Compensation: Traditional action recognition methods without explicit camera motion compensation are not working well. STIP, Cuboids and DT all performed poorly, in comparison to state-of-the-art method. Our first experiment is adding motion compensation. IDT with our head motion estimation and motion features significantly improves the results on all datasets (fourth row in Table 5.1), except for GTEA Gaze with 40 classes. This is due to imbalanced data as we examine the confusion matrix. While we expect better results by removing the camera motion, it is a bit surprising to find that IDT already provides comparable results with state-of-the-art methods.

Object vs. Motion: We proceed by supplementing object features in IDT. We compare three different settings as shown in the second group of Table 5.1: (1) IDT with motion features (M) along the trajectories as baseline; (2) IDT with object features (O) along the trajectories; (2) IDT with both object and motion features (O+M) . Even with simple object features, the results are surprisingly well, outperforming all previous states-of-the-art and motion features by a large margin. The results justify that object cues are crucial in understanding egocentric actions.

We notice that trajectories given by IDT provide rough location of foreground objects. Extracting object features along these trajectories is equivalent to extracting features on foreground moving objects, which is similar to [75]. Our object features encode which object the first person is interacting with, and thus is effective to recognize egocentric actions. Combining object and motion features (O+M), however, leads to marginal improvements, in comparison to using only object features.

Egocentric Cues: We further test egocentric features (E) in our method. These features are obtained by encoding the first-person's head motion and hand movements. Using only

egocentric features, we achieve a performance comparable to Cuboids. We then combine egocentric features with motion and object features (O+M+E), and only observe slight improvement over all datasets. Directly encoding egocentric cues is not effective. Head motion is less discriminative for fine-grain actions. For example, taking a slice of bread and taking a peanut butter jam can have very similar head motion. Moreover, hand movement is already encoded by the local motion features.

Trajectory Selection: In addition, we select descriptors based on their trajectories using manipulation or gaze point (O+M+E+G/H). We simply keep all the trajectories within the vicinity of an "attention" point, defined by a circle of radius r. The radius is defined by the minimum of the $2D$ distances between each point on the trajectory and the attention point in the corresponding frame. We vary the radius of the local region, plot the classification accuracy on all datasets in Figure 5.2 and report the best results in Table 5.1 (third group). We observe peaks along the curves. With a small region of radius equal to 60 pixels, roughly occupying 20% of the image area, our method is able to achieve a consistent performance boost from 2% to 16% over all datasets. This strategy is also very efficient as much less descriptors are encoded. The result indicates that "attention" points, e.g. gaze or manipulation points, provide a strong prior of where an action occurs.

In GTEA Gaze, the performance gap between manipulation points and gaze points is large. Again, we find that this result is dominated by categories with a few instances. In GTEA Gaze+, this gap is small. In fact, manipulation point has shown to be effective for gaze prediction [11]. While current evidence can not support the replacement of gaze points, we confirm that the concept of manipulation point is a powerful tool for egocentric action recognition. We also notice a plateau around the peaks of r, which suggests that our method is relatively robust to the measurement error of manipulation or gaze points.

Object vs. Motion Revisited: We further analyze which cue is more important with the selected descriptors. We benchmark object and motion features with the best radius (O/M+E+H/G) in the fourth group of Table 5.1. Constraining the features within a salient

Figure 5.2: Sensitivity study for action recognition. We encode features within radius r (pixels) around either a manipulation point (red) or a gaze point (green) for first person action recognition. We plot the recognition accuracy against the region size. The baseline accuracy at $r = 200$ is given by encoding all local descriptors within the video. Our trajectory selection improves the performance by choosing local descriptors relevant to FPV actions.

region improves the baseline performance of encoding object or motion features over the whole video. Moreover, object and motion features are complementary towards the final performance, except on GTEA, where object features are clear winners. This is largely due to the fact that GTEA used the same object instances in all actions under an ideal illumination.

Confusion Matrix: Our final results with O+M+E+H/G outperform previous results by a large margin. We improve the performance by 27.0% in GTEA, 13.9% in GTEA Gaze and 10.7% in GTEA Gaze+, in comparison to the state-of-the-art [76, 10, 137]. However, as we discussed in the beginning of the section, the single average class accuracy is not proper

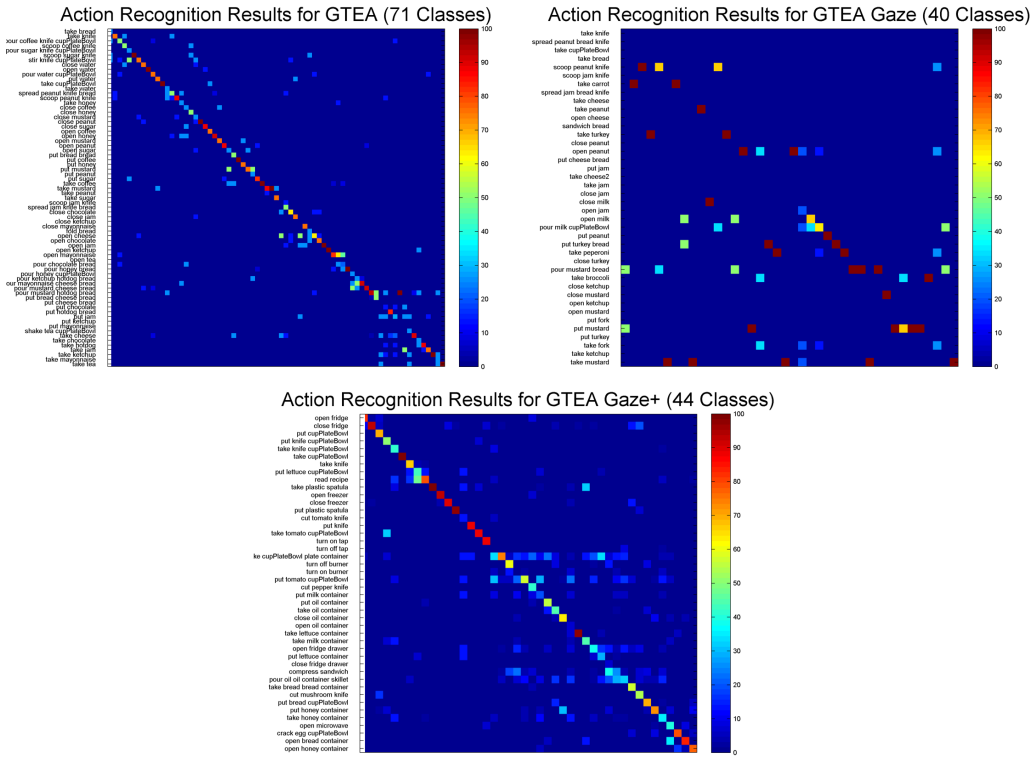

Figure 5.3: Confusion matrix of our method (O+M+E+H) on three datasets. Action categories are sorted based on decreasing number of instances. Our results are centered at the diagonal on GTEA and GTEA Gaze+. Our method achieves a performance boost of 27.0% in GTEA, 13.9% in GTEA Gaze and 10.7% in GTEA Gaze+ over the state-of-the-art methods [76, 10, 137].

for imbalanced data. To fully understand the results, we sort the action categories based on decreasing number of instances, and report confusion matrix using the combination (O+M+E+H) on all three datasets in Figure 5.3. Our method is able to get most categories correctly, except on GTEA Gaze. The result on GTEA Gaze is worse due to the mixture of low video quality, imbalanced samples and insufficient training data.

Best Practice for Dense Trajectories

Base on our experimental results, we recommend the combination of O+M+E+H for egocentric action recognition. We summarize and briefly explain the best practices for FPV action recognition using DT.

- Motion compensation is important. It leads to more reliable motion features, as well

as identifying foreground regions for meaningful object features.

- Object cues are of crucial importance in egocentric actions. The information of what object is been used greatly helps the performance of action recognition

- Using an "attention" point (manipulation/gaze point) to guide feature encoding works surprisingly well. A manipulation point derived from hand shape serves as a good approximation to the actual gaze point.

5.3 FPV Action Recognition using Deep Models

After the publication of our work, significant progress has been made for developing deep models for video classification. These models have shown to outperform traditional local descriptors on major action recognition benchmarks [7, 39, 41]. Several recent work have also explored deep networks for FPV action recognition [79, 81]. We want to understand whether our previous results and findings still hold for deep models, as these models greatly differs from the combination of hand crafted features and shallow classifiers. This has thus motivated us to extend our previous study by using deep models.

Deep models can learn directly from the input data, without the need of hand-crafted features. This property, however, makes it notoriously hard to interpret the results of deep models. As a remedy, our work leverages a explicit separation of input representations. More specifically, we represent object, motion and egocentric cues as individual streams in our network. This architecture thus allows us to study the importance of each cue for FPV actions. Thus, we present a multi-stream deep networks for encoding object, motion and egocentric features, and study how egocentric cues helps deep models for FPV actions. Figure 5.4 provides an overview of our approach.

Our study is in the same spirit as our previous work. We show that many of our previous findings can extend to deep models, yet careful thought should be given when training these models. For example, we demonstrate once again that egocentric cues can help to

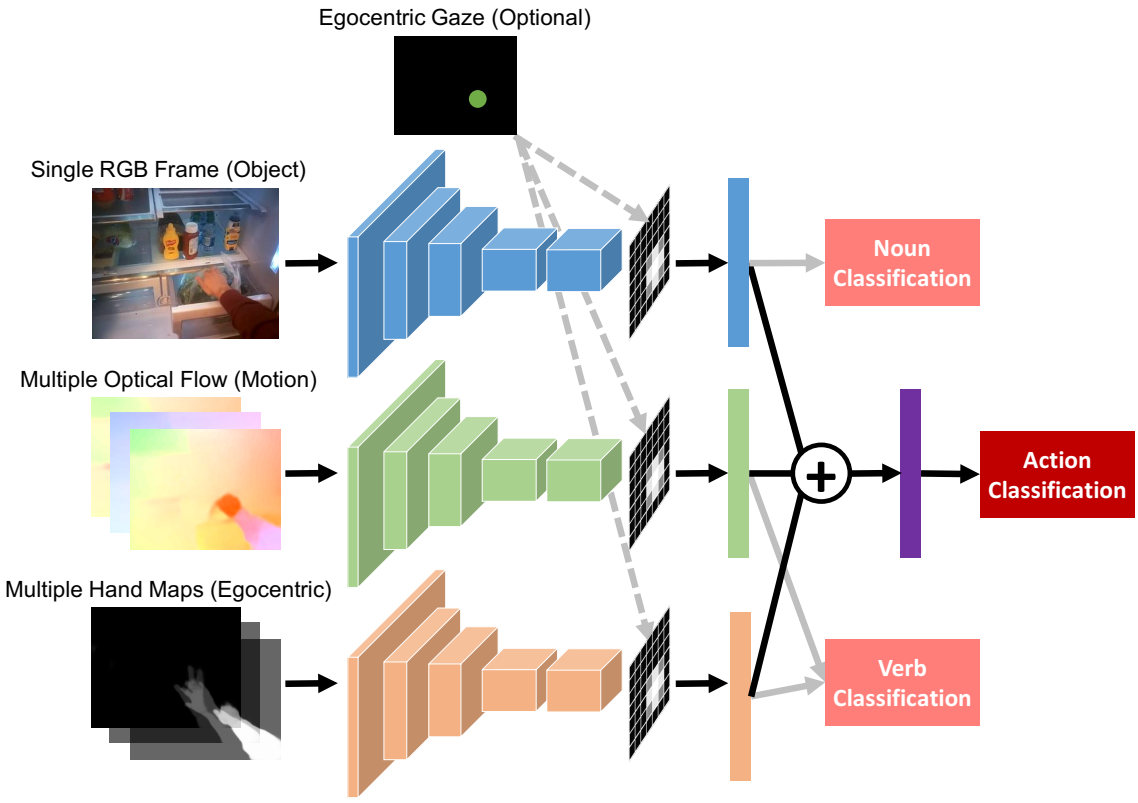

Figure 5.4: Multi-stream networks for FPV action recognition. Our full model has three streams. Motion and object streams remain the same as [39]. We add a separate egocentric stream to encode egocentric hand trajectories over time. Our model also incorporates a decomposed loss function as [81], and contains an optional attention mechanism for pooling features using egocentric gaze.

improve action recognition accuracy when using deep models. When not using gaze, our method achieves a performance that is on par with the state-of-the-art method [81], which requires additional annotation of objects. When using gaze, our method slightly outperforms previous best results.

This section is organized as follows. In Section 5.3.1, we first describe how we encode and combine object, motion and egocentric cues in deep models. In Section 5.3.2, We introduce our multi-stream networks, and discuss its implementation details and training schemes. Finally, in Section 5.3.3, we present experimental results and our findings.

5.3.1 Motion, Object and Egocentric Cues in Deep Models

We now present how to encode motion, object and egocentric cues in deep models. Our key idea is to use multi-stream networks to encode each cue using a separate steam. This strategy have been shown successful for action recognition [7].

Two-Stream Networks for Motion and Object Cues

Two-Stream Networks [7] provide a effective and simple architecture for action recognition. The key innovation of this method is to explicitly represent the input video by a spatial stream and a temporal stream. The spatial stream takes a single frame as its input and thus encode the appearance of the video, i.e. object cues. The temporal stream takes multiple optical flow as input and thus capture the motion cues. The output of the two streams are further fused at the end when classifying a video clip. It thus naturally separates motion and object cues. For example, we can benchmark impact of motion cue by the performance of a single temporal network.

We use the latest variant of Two-Stream Networks for FPV action, called Temporal Segment Networks (TSN) [39]. TSN further incorporate long-term temporal dynamics by sampling K snippets from the video and combining their outputs. This is done via a end-to-end learning manner. More specifically, a video clip is first divided into K temporal parts, and each time K frame sequences are individually sampled from each part. And the final score for the video is the average of K snippets. The idea is similar to temporal pyramid pooling [35] by adding redundancy into the video representation, yet did not fully capture the temporal ordering of the frames. The method produces the state-of-the-art results on major datasets.

Egocentric Cues Meet Deep Networks

We extract egocentric cues as previous section. These cues include head motion (homography), hand confidence map and 2D gaze points given by eye tracker. While two-stream

Networks provide a way for encoding motion and object cues, it is not clear how we encode egocentric cues under the same framework. This is because that egocentric cues are fundamentally different from video frames or optical flow. We present several techniques for incorporating egocentric cues with two-stream networks.

Motion Compensation Similar to our previous work, we can compensate the optical flow using head motion. This is done by subtracting the background motion induced by homography from the optical flow field.

Egocentric Stream We propose to directly feed the hand confidence map into a separate stream of networks. A hand confidence map has the same format of a gray scale image. We thus have the option of either using a single frame of hands, like the spatial stream, or multiple frames of hands, as the temporal stream. However, we found neither representations desirable for FPV actions. On one side, a single frame of hands is highly ambiguous. For example, the only way to distinguish between the actions of "take" and "put" is to know the temporal ordering of hand masks. On the other side, using multiple concatenated hand frames is highly redundant, as the hands change very little between frames. Moreover, there is no pre-trained model available for this new stream.

Therefore, we proposed a simple representation—Hand History Image to encode the temporal order of binary hand masks. We first threshold the hand confidence map to obtain a sequence of binary masks $D(x, y, t)$. We then construct the history image $H(x, y, t)$ for each time stamp t by

$$H_\tau(x,y,t) = \begin{cases} \tau & if D(x,y,t) = 1 \\ max(0, H_\tau(x,y,t-1) - o_t) & otherwise \end{cases} \tag{5.1}$$

The result is a scalar-valued image where more recently hand pixels are brighter. τ defines the max value of the image and is usually set to 255. o_t controls the decay of the hand pixels. Examples are presented in Figure 5.5. This representation is similar to the classic motion history image by Bobick and Davis [145]. Similar ideas have recently been explored

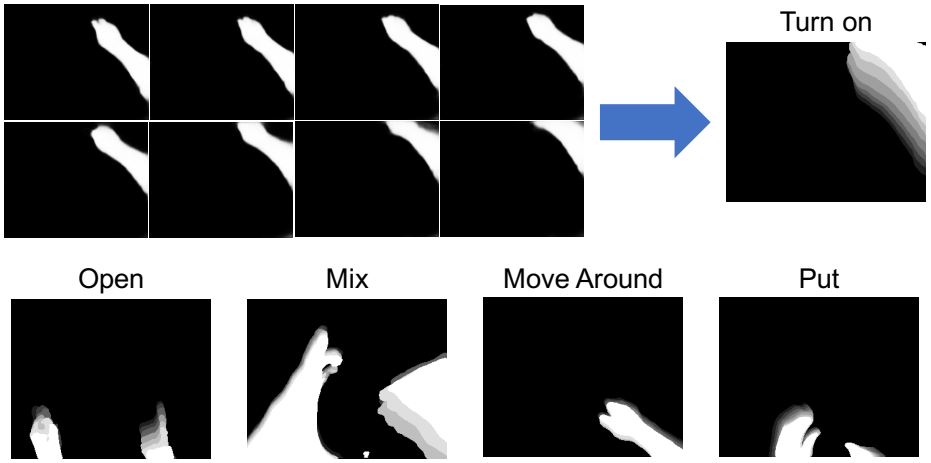

Figure 5.5: Examples of hand history image. Top row: a sequence of 8 hand confidence map and their hand history image. Bottom row: more examples of hand history image of different actions. These images can be used to distinguish the gross motion in a FPV action (e.g. verbs).

for deep models, for example by learning frame raking functions [146] or by creating 2D visual snapshots of video frames [147].

Attention Mechanism We further propose to incorporate egocentric gaze (g_x, g_y) into the network by using a soft attention mechanism. For a given stream, let $f_k(x, y)$ represent the activation of unit k in the last convolutional layer at location (x, y). We further assume a global average pooling[1] is performed to form the final feature vector $F_k = \sum_{x,y} = f_x(x, y)$. Average pooling is used in most recent network architectures, such as GoogLeNet [148] or ResNet [149]. F_k is further used as input to a fully connected or softmax layer. We replace the average pooling with weighted pooling by

$$\hat{F}_k = \sum_{x,y} \omega_{x,y} f(x, y) \quad where \quad \omega_{x,y} = \exp\left(\frac{\|x - g_x\|^2 + \|y - g_y\|^2}{\sigma^2}\right) \tag{5.2}$$

where the weights $\omega_{x,y}$ of location (x, y) is determined by its distance to current gaze point. σ determines the behavior of the attention mechanism. When σ is small, this is equivalent to using features around the gaze point as our previous work. When σ is large, it degenerates

[1] An attention mechanism for max pooling is possible, yet outside the scope of this document.

to average pooling. We set $\sigma = 1.0$ for our experiments. It is possible to joint learning the gaze and actions during training and thus does not require gaze for testing. We leave this as our future work.

5.3.2 Multi-Stream Networks for FPV Actions

We present our multi-stream networks for FPV action recognition, discuss the design of the loss functions, and present our training scheme of the network.

Multi-Stream Networks

Our multi-stream networks have three streams as shown in Figure 5.4. The spatial (object) and temporal (motion) streams remain the same as TSN. The hand (egocentric) stream take a gray scale hand history image as input. These three streams are fused at the last fully connected layer by adding all features together. Although more complicated fusion scheme exist [150], we found this simple sum provide satisfactory results[2]. Our architecture can optionally take the egocentric gaze as input. When available, these gaze points are used for pooling features at the last convolutional layers.

Moreover, we further decompose the action labels into verb, noun and actions. This decomposition leads to a joint training scheme of the model, which is discussed in [81, 151] and has been shown to provide good results. Specifically, we attach four loss functions to our network. A sigmoid cross entropy loss of noun labels are attached to the spatial (object) stream. This is because that object labels are not mutually exclusive. For example, the action of "put cup (on) plate" have two nouns. Two softmax cross entropy losses of verb labels are attached to motion streams. And finally, a softmax cross entropy loss is attached to the fused output. This is because both verb and action labels are mutually exclusive. All loss weights are set to 1.0 except the final action loss (3.0).

[2]With InceptionV2, this is equal to sum over the last convolutional block.

Training the Network

Training a deep models requires extra effort on smaller dataset. We have to carefully balance over-fitting (as the number of samples are small) and under-fitting (as the model has too many parameters). We present two important techniques to improve the training.

Pre-training and Multi-Stage Training: We train our network by fine-tuning pre-trained models. We use ImageNet pre-trained models for spatial and egocentric stream, and UCF pre-trained models for temporal stream. Stochastic Gradient Descent (SGD) with momentum is used to train all models. Moreover, we adapt a multi-stage training process. In the first stage, we disable the fused loss and train each stream independently. In the second stage, we freeze the lower part of all networks and jointly fine-tune all streams with all four loss functions.

Unbalanced Sample: Imbalanced data is a known major challenge for visual recognition, and in particular for deep models. When training deep models using SGD and a small batch size, the network can easily forget the concepts of less frequent classes, e.g. classes with small number of samples. We explored class re-weighting as in previous section, and found that it produces unstable gradients when using a small batch size. For example, the minority class can have a weight more than 20 where the batch size is 16. A single sample from the minority class will thus overwhelm the gradients of the whole mini-batch. Our solution is a combination of data resampling, aggressive data augmentation and the decomposed loss function.

Data Resampling: We oversample the data at the beginning of the training. After over sampling, all classes have similar number of instances. When combined with data augmentation, this is different from simply duplicating data points. We will see a different version of the same video when we hit it the second time.

Data Augmentation: We apply four types of data augmentations, following the order of random color perturbation, random flip, random rotation and random cropping. For the same clip, we independently sample frames for each stream. Frames within the same clip of

the same stream share the same augmentation within a mini-batch. Note that even the same augmentation can be very different when it is applied to a different stream. While data augmentation are widely used for action recognition, their details are usually not documented. We thus describe their details here.

- Color Perturbation: We randomly generate a small number (-5%-5%)to adjust the contrast of each channel. For spatial stream, this leads to perturbation of colors. For temporal and egocentric stream, this adds noisy in the data.
- Random Flipping: We randomly flip the image horizontally. Note for flow, this also requires modifying flow values in horizontal direction.
- Random Rotation: We randomly rotate the input image (-30 to 30 degrees) to mimic head rotation. For RGB frames and hand masks, we crop the maximum inner rectangle after the rotation. For flow, we apply 2D rotation matrix to flow vectors directly.
- Random Cropping: We randomly sample a bounding box with aspect ratio in the range of 0.67-1.33 and covers at least half of the area of the frame. We then crop the region using the sampled bounding box and resize it to a fixed resolution (224×224 in our case). For flow, we scale the flow values based on the resizing factors. When gaze is available, we make sure that the box always contains the gaze point.

Decomposed Loss Function: Finally, our decomposed loss function also provides a partial remedy to unbalanced samples. Decoupling an action's constituent verb and noun can affect the deep model's ability to learn these visual concepts separately and to combine them to perform recognition. We have found this loss critical for good recognition performance, as also discovered in [81].

Implementation Details

We use InceptionV2 [152] for all our experiments. For the first stage, we use a batch size of 128 to train each stream independently with learning rate 1e-4, momentum 0.9 and weight

decay 5e-4. Similar to TSN, we use a high dropout rate of 0.7 for all networks. We also apply partial batch norm, and only tune the first batch norm layer [152] of each network[3]. For the second stage, we freeze all layers below the fourth convolutional block (and thus no batch norm layer is trained) and use a batch size of 16 with decreased learning rate of 1e-5 and the same momentum and weight decay as the first stage. For spatial stream, we use a single RGB frame. For temporal stream, we use a concatenation of 5 flows. For egocentric stream, we use the hand history image from 8 images and set o_t to 30.

For testing, we follow the same pipeline as TSN by densely sampling 12 snippets for each input video, passing them through our model and averaging their output scores. For each snippets, we use the central crop and test on both the original video and its horizontally mirrored version. For video clips that does not have enough frames, we simply duplicate the snippets. This "ensembled" testing scheme helps to improve the recognition accuracy.

5.3.3 Experiments and Results

We describe our experiments and present the results. Not surprisingly, many of our previous findings using local descriptors still hold for deep models.

Dataset and Baselines

For all our experiments here, we use GTEA Gaze+, as other datasets does not provides enough samples for training deep models. Similar to our previous work, we resize all videos to 320×240 at 24Hz and follow the same evaluation criteria. We leave all experiments on the newly proposed Extended GTEA Gaze+ dataset as future work.

We compare our method to a set of baselines, including (1) our previous method that combines egocentric features with DT; (2) a modified version of two-stream network by using the decomposed loss function from [81]; (3) a modified version of TSN by using the decomposed loss function; (4) state-of-the-art method from [81], which uses additional

[3]We found it important to reduce the decay of exponential mean average in batch norm to e.g. 0.9. This modification allows batch norm to aggregate statistics much faster on a small dataset.

Table 5.2: Results of our multi-stream networks for FPV action recognition. Our main results include an ablation study and a comparison to baseline methods. Our best performing method incorporates motion compensation, egocentric stream and attention mechanism, and slightly outperforms the previous best results in [81], where additional object annotations are required for training.

	Methods	Accuracy
Baselines	Ego + DT	60.5
	Two-Stream [7]	58.8
	TSN [39]	62.4
	Ma et al. [81]	66.4
Ablation Study	+ Motion Compensation	63.8
	+ Multi-Stream (hand)	66.3
	+ Egocentric Gaze	67.0

annotations of object locations for training.

Results

Our results consist of two main parts: (1) an ablation study that adds one component at a time; and a comparison to baseline methods.

Ablation Study: To better understand how egocentric cues contribute to the performance of deep models, we run an ablation study of our method. We start with the vanilla TSN with modified loss functions, and gradually add motion compensation, egocentric stream and attention mechanism. The results are shown in Table 5.2. The vanilla TSN achieves an average mean class accuracy of 62.4% when combined with our training scheme and loss functions, already outperforming our previous best results at 60.5% using DT.

Adding motion compensation only slightly improves the accuracy by 1.4%, much less than the effect of motion compensation when using DT. It is possible that the deep networks can learn to remove the background motion. A similar argument is also discussed in [81]. Moreover, our multi-stream networks with egocentric hands, achieved an accuracy of 66.3, another 2.5% improvement. This improvement demonstrates the power of using a separate egocentric stream. Finally, our full model with attention mechanism offers another small boost of 0.7%. This confirms the observation in our previous work: egocentric gaze offers

marginal improvement for FPV actions when other egocentric features are available.

Comparison to Baselines: We further compare our results to baseline methods. Our previous best results using DT reached an accuracy of 60.5%. The modified two-stream network is at 58.8% and TSN is at 62.4%. These numbers suggest that deep models even without egocentric features, can approximate or outperform our previous work. Our best result without gaze is 66.3%, which is on par with the-state-of-art method [81] 66.4%. Unlike [81], we did not use additional object annotations for training. Our final model is slightly better than [81]. However, the gap 0.6% is not significant enough to draw meaningful conclusions. The important caveat is that adding egocentric cues improve the performance of TSN by a large gap of 4.6%.

Our results confirm our previous findings that egocentric cues offer complimentary information about FPV actions other than motion and object cues. Therefore, incorporating egocentric cues as an embodied representation can help to greatly improve the performance of FPV action recognition.

5.4 Conclusion

In this chapter, we demonstrate for the first time that egocentric cues can be incorporated with motion and object features to improve FPV action recognition. We establish rigorous benchmark baseline, and provide an extensive study of how object, motion and egocentric cues contribute to egocentric action recognition. Our method offers significant performance boost in major benchmarks. We explored two different video representations, and identified several key components for performance: motion compensation, object features over foreground regions, encoding egocentric hand into a separate channel, and the usage of an attention point to guide feature extraction. Our results suggest that first person visual cues is important for understanding FPV actions.

Our findings, derived from a large set of experiments, can be summarized into three parts: (1) Motion cues, with an explicit model of camera movement, can provide compa-

rable results with the state-of-the-art methods that use object-centric features. This result is surprising and challenges the prevailing view that motion features are less reliable in egocentric videos. (2) Object cues, even simple visual features, when combined with foreground regions, can significantly improve the performance in object related actions. This supports the argument of object-centric representations. (3) Egocentric cues, when combined with motion and object cues, can provide a further large increase in performance. The performance gap indicates that mid-level egocentric cues are crucial for FPV actions. We also discuss issues on implementation details and existing benchmarks.

It is worth noting that our action recognition is still limited to object manipulation tasks. And there is a vast range of our daily actions and activities beyond object manipulations. The opportunity of studying FPV actions and activities in our daily life, poses a grand challenge in computer vision. What the person is doing? When, where, why and how he or she is doing it? These open questions will motivate my future research agenda.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

First person videos have became a major source of video data. My thesis work on FPV, focuses on the development of novel embodied representations for FPV action recognition. I have developed dataset for FPV actions, and identified first person visual cues as an embodied representation. This representation is further used for my work on gaze estimation and action recognition. I plan to further advance my research by addressing FPV action and activity recognition in the wild, and exploring the applications of FPV in mobile health. In this final chapter, I summarize the contributions and discuss my future research agenda.

6.1 Conclusions

Here I revisit my thesis statement and link my contributions to the statement.

Thesis Statement: First person visual cues are unique properties of first person videos, which provide an embodied representation for estimating attention and recognizing actions.

My dissertation consists of three interrelated pieces: (1) FPV datasets and First Person Visual Cues; (2) FPV Gaze Estimation; and (3) FPV Action Recognition. I re-organized my key contributions of these pieces to support my thesis statement.

- My work on FPV dataset established the largest and most comprehensive benchmark for FPF hand segmentation, gaze estimation and action recognition.
- This dataset together with my study provided the first systematic analysis of the rich set of first person visual cues.
- These cues are further used as an embodied representation for both gaze estimation and action recognition in FPV.

- My work on egocentric gaze prediction showed for the first time that a subjects gaze can be estimated using first person visual cues, without using an eye tracker.

- My work on action recognition showed for the first time that first person visual cues can be used to significantly improve action recognition performance in FPV.

My work also extended to object segmentation [11, 73] and video summarization [63] using gaze. Independent of my thesis work, I have also explored detecting eye contact [153] and extracting physiological signals [58] from first person videos.

6.2 Future Work and Open Questions

A major limitation of my dissertation work is the narrowed scope of understanding actions and activities in object manipulation tasks. Going beyond object manipulation tasks, my long-term research goal is to solve the grand challenge of first person activity recognition in the wild, centered around the key problems of what the camera wearer is doing, and when, where, why and how he or she is doing it. I'd like to advance this research agenda by further developing computer vision methods for understanding FPV action and activities. In addition, I am interested in the applications of FPV in the field of Mobile Health.

6.2.1 FPV Action and Activities in the Wild

I will create and demonstrate the first comprehensive and accurate method for recognizing and predicting activities in FPV. Activity recognition is widely viewed as the "next frontier" of computer vision. My work will demonstrate the power and effectiveness of a first person embodied approach to activity analysis from video. A key step is the creation of a large-scale first person video dataset of naturalistic activities. I have already laid the foundations for this work at Georgia Tech, via a pilot study to collect naturalistic first person video from the Atlanta population. To achieve these goals requires more than just novel computer vision technology. I will leverage participant-centered design to construct usable mobile

hardware/software systems, and develop novel protocols and infrastructures for privacy preserving data collection and annotation at scale.

I will develop novel representations and methods for first person activity recognition. In particular, I will leverage deep learning to develop embodied models of activities. Two key research questions are: How are our activities shaped by our visual context of objects and scenes? And how is our perception of objects and scenes informed by our activities? Embodiment enables cross-modal learning of what we see and how we act. I am also interested in learning from rich linguistic descriptions of activities. They will provide an important opportunity for open-dictionary recognition and grounding of activities. I will develop methods to decompose an activity into key components of actions and further associate actions with their linguistic tokens. An additional research direction is to identify and model patterns of activities for individuals that make it possible to predict their current and future agenda.

My research focus on analyzing first person videos in the wild will be a step toward "open world" computer vision. FPV provides a means to describe the "long tailed" distribution that we believe characterizes each individual's visual experiencea small set of visual stimuli and events happens frequently, and a huge set occurs with low probability but collectively makes up a critically-important fraction. My research will address several key questions: How does the distribution of the activities and objects that define visual experience differ between individuals and groups? How can we design methods to recognize objects, scenes and activities that occur rarely? Developing solutions to these questions will advance core problems in computer vision, including video representation, object recognition and activity understanding.

I envision a system that can reliably recognize common visual concepts, such as objects, scenes and activities, in streaming first person video of our daily life. The system will thus provide the critical ability of continuous sensing our visual world, and further enable the opportunity of a ubiquitous all-encompassing personal assistanta very first step toward

assistive robotics.

6.2.2 FPV for Mobile Health

There is currently an exciting opportunity to develop FPV for Mobile Health. Lifestyle behaviors are consistently associated with morbidity and mortality rates for a number of chronic conditions. The ability to provide objective measures of a person's behavior and the contexts in which it occurs, can enable a new understanding of key health problems and support novel treatments. In this regard, cameras can complement traditional wearable sensors such as accelerometers. I have already explored measuring physiological and social signals [58, 153] from first person videos, connecting the field of FPV to Mobile Health.

I plan to further develop FPV for Mobile Health by integrating FPV with other mobile sensing modalities, such as inertial measurement and electrocardiogram. It will thus allow us to acquire an accurate estimate of the states of an individual, correlate these measurements with their contextual aspects, and identify environmental risk factors for diseases.

6.2.3 Open Questions

To summarize, I will develop novel datasets, methods and problems for the grand challenge of first person activity recognition. There are several important open questions.

- How can we create a naturalistic large scale dataset of first person daily activities?
- How can we develop deep learning methods for first person activity recognition? How can we incorporate embodied perception? How can we accurately model infrequent activities?
- What is the best way to integrate wearable cameras with other sensors for Mobile Health? What are the unique merits of FPV for measuring a person's behavior?

I am committed to march further along these directions. And I am passionate about working toward a world where wearable visual computing becomes a ubiquitous facet of life.

REFERENCES

[1] P. X. Nguyen, G. Rogez, C. C. Fowlkes, and D. Ramanan, "The open world of micro-videos," *CoRR*, vol. abs/1603.09439, 2016.

[2] D. Marr and A Vision, "A computational investigation into the human representation and processing of visual information," *WH San Francisco: Freeman and Company*, vol. 1, no. 2, 1982.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[8] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: a dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2011, pp. 2556–2563.

[10] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, 2012.

[11] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *ICCV*, 2013.

[12] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *CVPR*, 2015, pp. 287–295.

[13] A. Clark, *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press, 2008.

[14] P. Thagard, *Mind: Introduction to cognitive science*. MIT press Cambridge, MA, 1996, vol. 4.

[15] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[16] D. H. Ballard, "Animate vision," *Artificial intelligence*, vol. 48, no. 1, pp. 57–86, 1991.

[17] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International journal of computer vision*, vol. 1, no. 4, pp. 333–356, 1988.

[18] J. M. Findlay and I. D. Gilchrist, *Active vision: The psychology of looking and seeing*, 37. Oxford University Press, 2003.

[19] A. L. Yuille and A. Blake, *Active vision*. MIT Press, 1992.

[20] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, pp. 3559 –3565, 2001.

[21] D. H. Ballard, M. M. Hayhoe, P. K. Pook, and R. P. Rao, "Deictic codes for the embodiment of cognition," *Behavioral and Brain Sciences*, vol. 20, no. 04, pp. 723–742, 1997.

[22] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, pp. 266–277, 3 2001.

[23] C. Yu, D. H. Ballard, and R. N. Aslin, "The role of embodied intention in early lexical acquisition," *Cognitive Science*, vol. 29, no. 6, pp. 961–1005, 2005.

[24] C. Yu and D. H. Ballard, "Learning to recognize human action sequences," in *Development and Learning, 2002. Proceedings. The 2nd International Conference on*, IEEE, 2002, pp. 28–33.

[25] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 6, no. 03, pp. 337–359, 2009.

[26] J. Aggarwal and M. Ryoo, "Human activity analysis: a review," *ACM Comput. Surv.*, vol. 43, no. 3, 16:1–16:43, Apr. 2011.

[27] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[28] L. W. Campbell and A. E. Bobick, "Recognition of human body motion using phase space constraints," in *ICCV*, IEEE, 1995, pp. 624–630.

[29] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," in *Computer Vision, 1998. Sixth International Conference on*, IEEE, 1998, pp. 120–127.

[30] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer vision and image understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[31] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, IEEE, 2007, pp. 1–8.

[32] C. Wang, Y. Wang, and A. Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 915–922.

[33] B. Z. Yao, B. X. Nie, Z. Liu, and S.-C. Zhu, "Animated pose templates for modeling and detecting human actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 436–452, 2014.

[34] I. Laptev, "On space-time interest points," *International Journal of Compuer Vision*, vol. 64, no. 2-3, pp. 107–123, Sep. 2005.

[35] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[37] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006, ISBN: 978-3-540-33834-5.

[38] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005.

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *ECCV*, 2016.

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[42] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[43] G. Cheron, I. Laptev, and C. Schmid, "P-cnn: pose-based cnn features for action recognition," in *ICCV*, 2015.

[44] A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool, "Does human action recognition benefit from pose estimation?.," in *BMVC*, vol. 3, 2011, p. 6.

[45] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *CVPR*, 2008.

[46] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *CVPR*, 2012.

[47] Y. Tian, R. sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *CVPR*, 2013.

[48] A. Jain, A. Gupta, M. Rodriguez, and L. Davis, "Representing videos using mid-level discriminative patches," in *CVPR*, 2013.

[49] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *ECCV*, 2012.

[50] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *CVPR*, 2012.

[51] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[52] C. Li and K. M. Kitani, "Model recommendation with virtual probes for egocentric hand detection," in *ICCV*, 2013.

[53] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani, "How do we use our hands? discovering a diverse set of common grasps," in *CVPR*, 2015, pp. 666–675.

[54] G. Rogez, J. S. Supancic III, and D. Ramanan, "Understanding everyday hands in action from rgb-d images," in *ICCV*, 2015.

[55] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.

[56] G. Rogez, J. S. Supancic, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4325–4333.

[57] H. Jiang and K. Grauman, "Seeing invisible poses: estimating 3d body pose from egocentric video," in *CVPR*, 2017.

[58] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard, "Bioglass: physiological parameter estimation using a head-mounted wearable device," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, IEEE, 2014, pp. 55–58.

[59] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *CVPR*, 2017.

[60] Y. Hoshen and S. Peleg, "An egocentric look at video photographer identity," in *CVPR*, 2016.

[61] Y. Poleg, C. Arora, and S. Peleg, "Head motion signatures from egocentric videos," in *ACCV*, 2014.

[62] R. Yonetani, K. M. Kitani, and Y. Sato, "Ego-surfing first-person videos," in *CVPR*, 2015.

[63] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *CVPR*, 2015.

[64] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 38–55, 2015.

[65] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "Egosampling: fast-forward and stereo for egocentric videos," in *CVPR*, 2015, pp. 4768–4776.

[66] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: a survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 5, pp. 744–760, 2015.

[67] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," in *Pacific-Rim Symposium on Image and Video Technology*, Springer, 2011, pp. 277–288.

[68] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization.," in *CVPR*, 2012.

[69] H. S. Park, E. Jain, and Y. Sheikh, "3D social saliency from head-mounted cameras.," in *NIPS*, 2012.

[70] H. S. Park and J. Shi, "Social saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[71] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *TPAMI*, vol. 35, no. 1, pp. 185–207, 2013.

[72] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[73] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014.

[74] E. H. Spriggs, F. De la Torre Frade, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *IEEE Workshop on Egocentric Vision, CVPR 2009*, 2009.

[75] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *ICCV*, 2011, pp. 407–414.

[76] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *CVPR*, 2013.

[77] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: a first-person perspective," in *CVPR*, 2012.

[78] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *CVPR*, 2016.

[79] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *CVPR*, 2016.

[80] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: detecting hands and recognizing activities in complex egocentric interactions," in *ICCV*, 2015, pp. 1949–1957.

[81] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *CVPR*, 2016.

[82] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *CVPR*, 2011.

[83] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in *ECCV*, 2016.

[84] M. S. Ryoo and L. Matthies, "First-person activity recognition: what are they doing to me?" In *CVPR*, 2013.

[85] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," in *Advances in Neural Information Processing Systems*, 2015, pp. 2845–2853.

[86] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations with unlabeled video," in *CVPR*, 2016.

[87] M. S. Ryoo, "Human activity prediction: early recognition of ongoing activities from streaming videos," in *ICCV*, 2011.

[88] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *CVPR*, 2012.

[89] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba, "Predicting motivations of actions by leveraging text," in *CVPR*, 2016.

[90] K. M. Kitani, B. D. Ziebart, J. A. D. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, 2012.

[91] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv preprint arXiv:1412.6604*, 2014.

[92] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *ICLR*, 2017.

[93] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances In Neural Information Processing Systems*, 2016.

[94] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," in *CVPR*, 2017.

[95] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: forecasting from static images using variational autoencoders," in *ECCV*, 2016.

[96] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *ICML*, 2016.

[97] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: video forecasting by generating pose futures," *arXiv preprint arXiv:1705.00053*, 2017.

[98] K. Schindler and L. Van Gool, "Action snippets: how many frames does human action recognition require?" In *CVPR*, 2008.

[99] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *TPAMI*, vol. PP, no. 99, pp. 1–1, 2014.

[100] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *ICCV*, 2011.

[101] D. Xie, S. Todorovic, and S.-C. Zhu, "Inferring dark matter and dark energy from videos," in *ICCV*, 2013.

[102] Y. Zhou and T. L. Berg, "Temporal perception and prediction in ego-centric video," in *ICCV*, 2015.

[103] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *CVPR*, 2016.

[104] M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, and J. Feng, "Deep future gaze: gaze anticipation on egocentric videos using adversarial networks," in *CVPR*, 2017.

[105] G. Crabtree, E. Kcs, and L. Trahey, "The energy-storage frontier: lithium-ion batteries and beyond," *MRS Bulletin*, vol. 40, no. 12, 10671078, 2015.

[106] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, IEEE, 2011, pp. 3281–3288.

[107] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, vol. 2006, 2006, 5th.

[108] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *IJCV*, pp. 1–21,

[109] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: crowdsourcing data collection for activity understanding," in *ECCV*, 2016.

[110] G. A. Sigurdsson, O. Russakovsky, and A. Gupta, "What actions are needed for understanding human actions in videos?" In *ICCV*, 2017.

[111] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Opensurfaces: a richly annotated catalog of surface appearance," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 111, 2013.

[112] B. G. Fabian Caba Heilbron Victor Escorcia and J. C. Niebles, "Activitynet: a large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[113] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[114] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," *Image analysis*, pp. 363–370, 2003.

[115] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," *Pattern Recognition*, pp. 214–223, 2007.

[116] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *ICCV*, 2013.

[117] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow," in *CVPR*, 2015.

[118] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: evolution of optical flow estimation with deep networks," in *CVPR*.

[119] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[120] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[122] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, no. 25 - 26, pp. 3559 –3565, 2001.

[123] T. Kanade and M. Hebert, "First-person vision," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2442 –2453, 2012.

[124] A. K. Mishra, Y. Aloimonos, L. Cheong, and A. Kassim, "Active visual segmentation," *IEEE TPAMI*, vol. 34, no. 2, pp. 639–653, 2012.

[125] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[126] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254 –1259, 1998.

[127] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[128] M. F. Land, "The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations," *Experimental Brain Research*, vol. 159, pp. 151–160, 2 2004.

[129] L. M.F., "Predictable eye-head coordination during driving.," *Nature*, vol. 359, no. 24, pp. 318–320, 1992.

[130] L. Ren and J. Crawford, "Coordinate transformations for hand-guided saccades," *Experimental Brain Research*, vol. 195, pp. 455–465, 3 2009.

[131] M. Nystrom and K. Holmqvist, "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data," *Behavior Research Methods*, vol. 42, no. 1, pp. 188–204, 2010.

[132] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643 –659, 2005.

[133] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: a bayesian framework for saliency using natural statistics.," *Journal Vision*, vol. 8, no. 7, pp. 32.1–20, 2008.

[134] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *NIPS*, 2008, pp. 681–688.

[135] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010, pp. 3241–3248.

[136] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollar, "Exploring weak stabilization for motion feature extraction," in *CVPR*, 2013.

[137] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[138] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *CVPR*, 2013.

[139] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[140] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *ICCV*, 2013.

[141] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: training and applications," *Technical Report AIM*, 1997.

[142] J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *ICCV*, 2011.

[143] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[144] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, IEEE, 2013, pp. 245–251.

[145] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[146] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *CVPR*, 2015.

[147] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *CVPR*, 2016.

[148] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[149] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[150] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[151] Y. C. Zhang, Y. Li, and J. M. Rehg, "First-person action decomposition and zero-shot learning," in *WACV*, 2017.

[152] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[153] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg, "Detecting bids for eye contact using a wearable camera," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, IEEE, vol. 1, 2015, pp. 1–8.